New Methods for Multivariate Distribution Forecasting

Yang Han

University of Sussex

A thesis submitted for the degree of

Doctor of Philosophy

April 2019

Acknowledgement

Doing a PhD has been described to me – by my supervisors and other advisers – through many similes. I remember when Carol argued that research is like baking bread, a hobby that I picked up during my time in the UK: It needs a lot of work and preparation, but also sufficient time to grow. As I am nearing the end of this unique journey, I would like to add to this analogy: Grinding and persistence may be needed but in the end it is all worth it, because you end up inventing your own *sliced bread*.

First of all, I want to express my gratitude to Carol Alexander for her dedication, guidance and insightful comments throughout the PhD. It was a great pleasure working alongside such a brilliant and passionate scholar. I like to think that the last few years have helped me to grow as a researcher and person. If this is indeed the case, it all is thanks to her.

My sincere thanks also goes to Michael Coulon for his constant support. His fresh perspectives coupled with his innate intelligence were a huge aid and comforting presence during the last years.

Last but not the least, I want to thank my parents as well as numerous friends and colleagues who made the time so enjoyable.

Declaration of original authorship

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced. This thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Munich, 22 April 2019
Yang Han

Abstract

This thesis contributes to the current literature in finance and economics by introducing new methods for forecasting and accuracy evaluation. First, we propose and develop a new multivariate distribution forecasting method. Second, we compare proper scoring rules through a discrimination measure.

Our Factor Quantile models are flexible semi-parametric models for multivariate distribution forecasting where conditional marginals have a common factor structure, their distributions are interpolated from conditional quantiles and the dependence structure is derived from a conditional copula. A version based on latent factors can be constructed using endogenous principal component analysis. We present a comprehensive comparison of Factor Quantile models with GARCH and copula models for forecasting different multivariate distributions which is the first extensive application of proper multivariate scoring rules for financial asset returns. Our empirical study employs daily USD exchange rates from 1999 – 2018; US interest rates from 1994 – 2018; and Bloomberg investable commodity indices from 1991 – 2018 with eight time series in each system, yielding almost 1 million predictions. Formal testing indicates favourable forecasting performance of Factor Quantile models, matching or exceeding the accuracy of more complicated GARCH models, which take at least six times longer to calibrate and may also exhibit difficulties with parameter optimisation even when the multivariate distribution has only few dimensions.

In a simulation study, we analyse the ability of multivariate proper scoring rules to determine the true data generating model. We apply a new discrimination measure to the energy score and different parameterizations of the variogram score. Then, we evaluate the performance of this metric in standard tests of superior predictive ability. Previous literature generally agrees that the ideal score depends on the data and models. However, our findings clearly identify the variogram score with $p = 1$ as the most successful score in all three data sets, largely irrespective of the choice for the data generating model.

Contents

1 Introduction **1**

2 Literature Review **9**
2.1 Distribution Forecast Evaluation . . . 12
2.2 Distribution Forecasting with Quantiles . . . 17
2.2.1 Lagged Forecasting Models . . . 19
2.2.2 Contemporaneous Forecasting Models . . . 24
2.2.3 Alternative Forecasting Models . . . 28

3 Theoretical Background **31**
3.1 Quantile Regression . . . 33
3.2 Principal Component Analysis . . . 38
3.3 Copulas . . . 42
3.4 Univariate GARCH Models . . . 48
3.5 Multivariate GARCH Models . . . 51

4 Factor Quantile Methodology **54**
4.1 Factor Quantile Regression . . . 57
4.2 A Simple Example on Stock Returns . . . 63
4.3 Factor Quantiles with Latent Factors . . . 69
4.3.1 Alpha Quantile Forecasts . . . 73
4.3.2 Bagging Quantile Forecasts . . . 77

5 Distribution Forecast Evaluation **83**
5.1 Formal Tests of Forecast Performance . . . 86
5.1.1 Diebold-Mariano Test . . . 88
5.1.2 Model Confidence Sets . . . 90
5.2 Proper Scoring Rules . . . 93
5.2.1 Continuous Ranked Probability Scores . . . 97
5.2.2 Energy Scores . . . 101
5.2.3 Variogram Scores . . . 103

6 Empirical Study **106**
6.1 Data Description . . . 110
6.2 Empirical Design . . . 117
6.3 Forecasting Accuracy Results . . . 124
6.3.1 Univariate Forecasting Accuracy . . . 126

6.3.2 Multivariate Forecasting Accuracy 140

7 Simulation Study **148**
7.1 Simulation Design 153
7.2 Simulation Results 157
7.2.1 Sample Mean Comparison 159
7.2.2 Error Rate of Scoring Rules 162
7.2.3 Discrimination Heuristic of Scoring Rules 166

8 Summary and Conclusions **170**

A Model Confidence Set Tables **A1**

B Table Extracts **A10**

List of Figures

3.1 Asymmetric penalty for quantile regression with various parameters . 35
3.2 Quantile regression hyperplane for CAPM on Apple stock returns . . 35
3.3 Percentage of variance explained by first principal components 40
3.4 Densities of popular copulas ($d = 2$) 46
3.5 Asymmetric response function of E-GARCH for various parameters . 49

4.1 Distribution estimates with varying quantile partitions (Apple) 64
4.2 Conditional distribution and density forecasts (Apple and P&G) . . . 66
4.3 Joint conditional density forecasts (Apple and P&G) 66
4.4 Joint conditional density forecasts (Apple and P&G) 67
4.5 Forecast with first principal factors (6 month interest rate) 73
4.6 Forecast with last principal factors (6 month interest rate) 76
4.7 Principal component densities (US interest rate changes) 78
4.8 Forecast with asymptotic bagging (6 month interest rate) 81

5.1 CRPS Schematic 98
5.2 Variogram observation of various orders 104

6.1 Daily returns / changes on all three data sets 112
6.2 Regimes for US interest rates 114
6.3 Regimes for exchange rates and commodity indices 114
6.4 Exchange rate outliers 115
6.5 Cumulative variance explained by the principal components 120
6.6 Convergence issues with GARCH models (sugar) 122
6.7 Comparison of calibration time (commodity index returns) 123
6.8 FQ-AL_{250}: Uniformly weighted CRPS test statistic 138
6.9 FQ-AB_{250}: Uniformly weighted CRPS test statistic 139

7.1 Average scores relative to score of DGP (USD exchange rates) 159
7.2 Density of differences between scores with DCC-GARCH as DGP . . 163
7.3 Error rates of scoring rules 164
7.4 Discrimination heuristic of scoring rules 167

List of Tables

4.1 Kolmogorov-Smirnov p-values of distribution comparison (Apple) . . 65
4.2 Information criteria for the copula fit (Apple and P&G) 67

5.1 Possible weights for CRPS 99

6.1 Sample for each data set 111
6.2 Summary statistics of the monthly returns / changes 113
6.3 Summary of models used in the empirical study 120
6.4 MCS p-values for FQ-AL: Uniformly weighted CRPS 128
6.5 MCS p-values for FQ-AB: Uniformly weighted CRPS 129
6.6 Summary of CRPS hypothesis tests 132
6.7 Comparison of univariate performance over time 134
6.8 Performance over different regimes of US interest rates 136
6.9 MCS p-values for FQ-AL: Multivariate scores 141
6.10 MCS p-values for FQ-AB: Multivariate scores 143
6.11 Comparison of multivariate performance over time 145

List of Abbreviations

CAPM	Capital Asset Pricing Model
CDF	Cumulative distribution function
CRPS	Continuous Ranked Probability Score
CCC-GARCH	Constant Conditional Correlation GARCH
DCC-GARCH	Dynamic Conditional Correlation GARCH
DGP	Data generating process
EDF	Empirical distribution function
ES	Energy Score
FQ-AB	Factor Quantile with last principal components
FQ-AL	Factor Quantile with asymptotic bagging
GARCH	Generalised Autoregressive Conditional Heteroscedasticity
MCS	Model Confidence Set
PCA	Principal Component Analysis
VS	Variogram Score

List of Mathematical Notations

$\mathcal{O}(f), o(f)$	Bachmann–Landau notation for the limit behavior relative to f
$\mathcal{U}(a, b)$	Continuous uniform distribution on $[a, b]$
$\det(\mathbf{X})$	Determinant of a matrix $\mathbf{X}$
$\operatorname{diag}(\mathbf{X})$	Diagonal matrix composed of the diagonal elements of matrix $\mathbf{X}$
$\operatorname{diag}(\mathbf{x})$	Diagonal matrix composed of the elements $\mathbf{x}$
$\operatorname{dom}(f)$	Domain of a function f
$\mathbb{1}\{P\}$	Iverson bracket, yielding 1 if proposition P is true and 0 otherwise
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\operatorname{ran}(f)$	Range of a function f
$\operatorname{rg}(x)$	Rank of x among some sample
$\operatorname{sgn}(x)$	Signum function for a real number x
t_ν	Student-t distribution with ν degrees of freedom
$\operatorname{tr}(\mathbf{X})$	Trace of square matrix $\mathbf{X}$

Chapter One

Introduction

Statistical forecasts of multivariate time series guide many interactions in finance and economics with the primary goal of obtaining information of some future random vector. Whatever will be will be, but it is arguably better to know how things will turn out beforehand. It comes to no surprise then, that many authors have requested statistical forecasts to be of a probabilistic nature early on (de Finetti, 1975; Dawid, 1984). Ideally, the entire distribution of multivariate time series should be studied since this yields the most comprehensive view.

However, despite this general agreement, traditional literature on financial and economic variables has focussed almost entirely on point forecasts, typically based on the mean, often with some measure of forecast uncertainty such as the variance about this mean. But this only represents a distribution forecast under simplifying assumptions such as normality and is often not sufficient to derive optimal recommended actions, especially when users have heterogeneous loss functions (Elliott and Timmermann, 2016).

The prevailing focus on point forecasts does not imply a complete absence of probabilistic forecasts. Most notably, the literature experienced a surge of probabilistic predictions at the start of the century with papers such as Palmer (2002) and Gneiting and Raftery (2005) in meteorology, Garratt et al. (2003) in macroeconomics as well as Timmermann (2000), Groen et al. (2013) and Duffie and Pan (1997) in finance and economics. Nevertheless, as we will see further in the literature review in Chapter 2, these examples do not extend fully to distributional forecasts and are limited to the univariate case. Predictions of the entire distribution function (rather

than focussing on specific parts) remain somewhat rare in general and this becomes even more pronounced when considering multivariate distribution forecasts.

Proliferation of univariate point forecasts may be partly attributed to the challenges encountered when one tries to extend methodologies to higher dimensions. With increasing dimension, parameter estimation and calibration become more complex and error-prone. This curse of dimensionality impedes applications of high-dimensional distribution forecasts and makes them cumbersome to use. In fact, even relatively simple concepts such as quantiles may not have a satisfying multivariate equivalent because there is no unique way to invert a multivariate distribution function and no inherent ordering in multiple dimensions.

Our motivation for Factor Quantile models is to circumvent these problems entirely by deriving a multivariate distribution forecast through a conditional copula on marginals generated from univariate factor model quantile regressions. The fundamental steps in the algorithm are easily understood in three stages:

Stage 1 For each dependent variable, we predict a range of conditional quantiles in $(0, 1)$ using univariate quantile regression on multiple common factors;

Stage 2 For a given realisation of common factors, we then estimate a conditional distribution for each dependent variable using shape-preserving interpolation;

Stage 3 Dependence between these conditional marginals is imposed by the choice of copula, thus generating a multivariate distribution where the marginals are derived from univariate conditional distributions estimated via factor model quantile regressions.

We introduce a latent version of our Factor Quantile model that takes advantage of the dimensionality of the forecasting problem through principal component analysis. This yields Factor Quantile specifications that are not reliant on any externally generated forecasts or require ex-ante selection of suitable predictors. Our algorithm is very fast and flexible, and because the quantile regressions are univariate it scales well as the number of variables increases. Further it is applicable as a general

multivariate distribution forecasting methodology to many data sets. In comparison, multivariate quantile regression approaches, such as those proposed by Chakraborty (2003) or Chavas (2018), require a vast data set and are much more computationally intensive.

Examining how predictive ability is conditional not only on the choice of model, but also on the sample data and/or the objectives for parameter estimation, Giacomini and White (2006), Machete (2013) and Elliott and Timmermann (2016) all argue that there is no single superior approach: the best model or method depends on the statistical properties of the data and the economic properties of the variable being forecast. Thus, we employ multivariate time series over exceptionally long time periods, focussing on the unique properties of three very different multivariate data sets. In each case, we draw different conclusions about the most accurate forecasting model.

Our empirical study employs daily USD-denominated exchange rates from 1999 – 2018; US interest rates from 1994 – 2018; and Bloomberg investable commodity indices from 1991 – 2018. With eight time series in each of three multivariate systems we have a total of over 96,000 out-of-sample observations and with daily rolling re-calibrations of 14 different multivariate models on each data set we obtain over 1.3 million distribution forecasts to be evaluated. The scale of this study sets it apart from previous work on time series forecasting. Several recent papers also introduce new time series models within our three data sets but these only generate point forecasts.[1]

To assess the performance of Factor Quantile models, we compare them with two multivariate distribution forecasting models that are often applied to systems of financial and economic variables: (i) asymmetric Student-t multivariate GARCH(1,1) models, and (ii) empirical marginals with Gaussian copulas. These have been selected as (i) the family of parametric models which best capture the salient properties

[1]For USD-denominated exchange rates see Kilian and Taylor (2003) and Greenaway-McGrevy et al. (2018); for the US interest rate term structure see Bali et al. (2009) and Almeida et al. (2017); and for commodity futures see Chen et al. (2014) and Zolotko and Okhrin (2014) – amongst many others.

of financial time series i.e. volatility clustering, skew and heavy tails, asymmetric response to shocks, and (ii) a copula which is amenable to the high-dimensional systems which also performs well in previous forecasting exercises (Patton, 2012, 2013). Of course, there are a plethora of models available but including further models would provide so much information as to detract from the clear messages of this study.

We use proper scoring rules to quantify the forecasting accuracy of all distribution forecasts. Suppose the data generation process is the distribution F. A scoring rule is proper if the expected score is minimized when the forecaster issues the probabilistic forecast F, rather than another distribution $G \neq F$, and it is strictly proper if this minimum is unique. Since the goal of probabilistic forecasting is to maximize sharpness of the distribution forecast, subject to calibration, proper scoring rules are particularly advantageous as they address both calibration and sharpness simultaneously (Winkler, 1996).[2] Also, as recommended by Gneiting et al. (2008) and Scheuerer and Hamill (2015) we utilize multiple univariate and multivariate proper scores, since the high degrees of freedom for the forecasts make it unlikely that a single score can serve all purposes.

Scoring rules are convenient because they summarize the forecasting performance into a single score that tests for calibration and sharpness. Although there are several scores for the assessment of univariate probabilistic forecasts, the rankings they provide generally coincide (Staël von Holstein, 1970; Winkler, 1971; Bickel, 2007) so that there are no conflicting conclusions. However, we find this not to be the case for multivariate scoring rules during our accuracy comparison. This may be attributed to the high degree of freedom which leads to a large loss of information during the encapsulation of the performance into a single score. We extend the parsimonious comparisons of multivariate scoring rules of Scheuerer and Hamill (2015) and Pinson and Tastu (2013) by conducting an extensive simulation analysis and we assess the

[2]Calibration is the statistical consistency between a distribution forecast and the observations while sharpness is the concentration of the forecast distribution. As such, they are similar in concept to unbiasedness and efficiency of statistical estimators.

ability of the proper scoring rules to differentiate between erroneous distributions and the true distribution given realistic conditions.

This doctoral thesis makes three primary contributions to the literature on multivariate distribution forecasting:

(i) First, we propose a new, semi-parametric model for estimating and forecasting multivariate distributions where marginals are derived from factor model quantile regressions (Koenker and Bassett, 1982) and the dependence structure is modelled using a conditional copula (Patton, 2006). It may be applied with any macroeconomic, fundamental or statistical factor model; each has the advantage that a dependence structure in the original, larger system is conditional on dependence between relatively few factors. Our latent Factor Quantile version uses principal components as factors and yields a general forecasting methodology that does not rely on any externally generated forecasts or ex-ante predictor selection.

(ii) Second, we present the first extensive financial application of proper multivariate scoring rules, previously developed in meteorology and other branches of atmospheric science (Jolliffe and Stephenson, 2003; Keune et al., 2014), to assess the accuracy of daily time series forecasts for three different systems: exchange rates, interest rates and commodity futures.[3] Only a few previous empirical applications of multivariate scoring rules can be found in the literature, and these are to weather ensemble forecasts. In fact, most prior research in empirical finance has limited forecast evaluation to certain quantiles, such as Value-at-Risk, typically of a univariate distribution. There are a few recent applications of proper scoring rules to financial or economic data, but these have been to point forecasts or univariate distributions over a single out-of-sample period – see Panagiotelis and Smith (2008), Hua and Manzan (2013), Ravazzolo

[3]Diks et al. (2010, 2014) evaluate the out-of-sample performance in their studies through the likelihood function. This corresponds to the application of the strictly proper multivariate logarithmic score. However, the logarithmic score has been criticised by some for its heavy penalty on low probability events and hence may not be suited for the forecast evaluation.

and Vahey (2014), Manzan (2015), Alexander et al. (2019) and Meligkotsidou et al. (2019).

(iii) Third, we study the ability of multivariate scoring rules to differentiate the true distribution against various misspecified forecasts in a realistic setting. For this, we conduct a simulation study that compares the energy score and various parameterisations of the variogram score over a long evaluation period. We show that the scoring rules differ significantly with respect to their discrimination ability and derive recommendations on their application in practical settings.

The remainder of this thesis is structured as follows. Chapter 2 motivates our research by critically surveying the related financial and econometric literature. We focus on previous studies on univariate distribution forecasting with quantile regression, multivariate forecasting models and out-of-sample forecast evaluation based on proper scoring rules.

All mathematical prerequisites for this thesis are described in Chapter 3, starting with a summary of the relevant aspects of quantile regression, principal component analysis and copulas. Subsequently, we discuss univariate and multivariate GARCH models that we use as benchmark in the empirical study to showcase the relative forecasting accuracy of our methodology.

The Factor Quantile methodology is introduced in Chapter 4. We start with a description of the general method that derives conditional marginal distributions non-parametrically by quantile regression and combines them to construct a conditional joint distribution with a parametric conditional copula. Similar to the forecasting model by Gaglianone and Lima (2012), this approach uses a contemporaneous linear model to translate externally generated point forecasts of common factors to a distribution forecast of dependent variables. Then, we focus on the case where latent factors are derived using endogenous principal component analysis and present two versions of our model that rely neither on the availability of appropriate macroeconomic or fundamental linear factor models nor externally generated forecasts. Examples on a simple bivariate case of two US stocks and on eight-dimensional US

interest rates motivate our recommendations for the choices to be made regarding quantile partition, interpolation method and the latent factor model applied to the conditional marginals.

We summarize the theoretical background of our forecast accuracy evaluation in Chapter 5 and justify our choices for the evaluation methodology. This starts with a discussion on the formal tests of forecast performance that have been used in literature to compare the predictive power of competing models. Then, we describe popular proper scoring rules which measure the accuracy of univariate and multivariate distribution forecasts and act as loss functions in the hypothesis tests.

Chapter 6 presents our empirical study where we compare two specifications of our semi-parametric model against standard econometric model classes for forecasting systems of exchange rates, the term structure of interest rates and commodity future indices. First, we test univariate distribution forecasts using the weighted continuous ranked probability score (CRPS) proposed by Gneiting and Ranjan (2011), which has the advantage of allowing different weight functions to assess accuracy in the lower tails, in the upper tails, in both tails, in the centre and in the entire distribution. Then, we apply the energy and variogram scores to measure the accuracy of multivariate distribution forecasts – see Gneiting et al. (2008) and Scheuerer and Hamill (2015). In each case we compare the relative accuracy of the entire set of distribution forecasts considered in our empirical study through the equivalence test and elimination rules of the Model Confidence Set (MCS) of Hansen et al. (2011);

We analyse the discrimination ability of multivariate proper scoring rules in Chapter 7. Our simulation study is designed such that the true future distribution is known and applies the same models as in our empirical evaluation chapter. Contrary to prior studies comparing scoring rules, we ensure a realistic setting with real data rather than limiting the discussion to simple parametric distributions as data generating processes. We show that the focus on propriety is not a sufficiently strict requirement for a good scoring rule and provide clarification in case different multivariate scores yield conflicting conclusions.

Chapter 8 concludes the dissertation by summarizing our results for our new semi-parametric Factor Quantile model and for our simulation study on proper scoring rules in a realistic setting.

All code in Python, MATLAB and R as well as all three data sets used in this thesis are available from the author on request. Further, a large number of additional figures and tables are accessible electronically in our supplementary materials.

Chapter Two

Literature Review

2.1 Distribution Forecast Evaluation 12
2.2 Distribution Forecasting with Quantiles 17
2.2.1 Lagged Forecasting Models 19
2.2.2 Contemporaneous Forecasting Models 24
2.2.3 Alternative Forecasting Models 28

The literature on forecasting financial and economic variables has advanced considerably during the last few years, beyond the traditional view of point forecasting, to focus on the fact that a forecast of a random variable is a distribution, by definition. Point forecasts are typically based on the mean, often with some measure of forecast uncertainty such as the variance about this mean. But this only represents a distribution forecast under simplifying assumptions such as normality and many studies demonstrate the need for a more comprehensive, probabilistic characterisation: Harvey and Siddique (2000), Dittmar (2002) and others show that third or even fourth moments explain cross-sectional variation in US stock returns; Amin and Kat (2003) employ the entire distribution of hedge fund returns to evaluate a manager's performance; and in portfolio optimization the whole multivariate distribution forecast for asset returns is required to calculate the investor's expected utility – see Birge (2007) and Resta (2012) for reviews. Distributional forecasts are especially relevant in situations where many users with heterogeneous loss functions rely on the prediction, since point forecasts are not sufficient to derive optimal actions to recommend in this scenario (Elliott and Timmermann, 2016).

The two most common econometric models for forecasting multivariate distributions of financial asset returns are generalised autoregressive conditional heteroscedasticity (GARCH) models (Bollerslev, 1990; Engle, 2002) and copula models (Patton, 2013). Of particular relevance to marginals generated by factor models, Patton (2006) extends the theory of copulas to allow for conditioning of variables, and Patton (2012) illustrates how conditional copulas are used for economic forecasting in the two-dimensional setting of small-cap and large-cap US equity indices. In the Bayesian forecasting literature, Markov Chain Monte Carlo (MCMC) is a popular estimation method for the posterior distribution but requires careful assessment of the convergence to avoid misleading inferences (Karlsson, 2013). However, numerous other models have been proposed in a voluminous strand of the literature which is critically surveyed by Elliott and Timmermann (2016). They emphasise that there is no single superior approach: the best method or model depends on the

statistical and economic properties of the variables concerned. Indeed, because of model misspecification and parameter estimation issues, better forecasts often result from combinations of different models.

2.1 Distribution Forecast Evaluation

Alongside the profusion of models for generating point or distribution forecasts, a prolific strand of theoretical research has focussed on developing methods for evaluating these forecasts. Elliott and Timmermann (2016) provide an overview of some elementary tests based on loss differentials, especially those introduced by Diebold and Mariano (1995) and Giacomini and White (2006). While Diebold and Mariano (1995) develop out-of-sample tests which compare errors of point forecasts, Giacomini and White (2006) extend these tests to multi-step point, interval or entire (univariate) distribution forecasts, and consider how predictive ability is conditional on the choice of data and/or objectives for parameter estimation. In settings with a large number of pairwise-comparisons, the model confidence set (MCS) of Hansen et al. (2011) may be particularly useful because this set contains all models for which forecasting accuracy cannot be distinguished at a specified confidence level. Unlike the Hansen (2005) tests for superior predictive ability (SPA), the MCS applies only simple hypotheses tests at each iteration and this facilitates its computation on large-scale data sets.

Concerning the loss function for the hypothesis tests, the standard approach is to quantify the accuracy of each prediction with a proper scoring rule – see Gneiting and Raftery (2007) for further discussion. For instance: Bao et al. (2007) advocate using the Kullback–Leibler information criterion which is derived from the logarithmic score, a proper rule that has been criticised by some for its heavy penalty on low probability events; Boero et al. (2011) find that ranked probability scores have better discriminatory power than logarithmic or quadratic scores; Gneiting and Raftery (2007) advocate using the continuous ranked probability score (CRPS); and Gneiting and Ranjan (2011) extend this to adopt the weighting approach of Amisano and Giacomini (2007) so that evaluation can be focused on a specific area of the distribution, such as a tail or the centre.[1] This has many advantages for forecasting

[1] Amisano and Giacomini (2007) compare density forecasts using a weighted likelihood ratio test, but this is not a proper scoring rule. With a proper scoring rule, forecasters gets the best

financial variables, where the accurate forecasting of tail risks is particularly important for risk management. In the case of multivariate forecasts, Gneiting and Raftery (2007) generalize the CRPS to the energy score, while Scheuerer and Hamill (2015) use the concept of variograms from geostatistics to derive a variogram score. Further, Dawid and Sebastiani (1999) introduce a scoring rule that is proper relative to the class of distributions with finite second moments and strictly proper if additionally the distributions are fully characterized by the first two moments.

Given the large number of scoring rules, conventional wisdom dictates to apply a suitable one for the application at hand. While it is generally agreed upon that only proper scoring rules quantify the accuracy of probabilistic forecasts adequately (Winkler, 1996; Gneiting and Ranjan, 2011), the question of which of the proper scores to use remains largely open (Gneiting and Raftery, 2007). This problem is especially relevant for multivariate evaluation, since the rankings of univariate scoring rules mostly coincide, which reduces the risk of conflicting conclusions (Staël von Holstein, 1970; Winkler, 1971; Bickel, 2007).[2]

Several studies analyse proper scoring rules analytically to derive recommendations for the choice of a suitable scoring rule for specific forecasting problems but those are restricted to an univariate setting and make strong assumptions. Machete (2013) compares how univariate scoring rules react to deviations between a forecast and the true but unknown distribution if this difference can be modelled as an odd function. However, his results do not yield sufficient guidance on the scoring rule selection apart from some generic suggestions and are further not valid for forecasts of non-symmetrical distributions which limits the applicability of their findings. Buja et al. (2005) apply a tailored approach to binary probability estimations and introduce a beta family of proper scoring rules that allows for a heuristic selection

score by forecasting their true beliefs. It is a strictly proper rule if it is not possible to get the same score by forecasting something else.

[2]Proper scoring rules are sometimes used to fit models, similar to maximum likelihood. In this branch of the literature, different univariate scoring rules may sometimes yield varying parameter estimates. However, this is not contradictory since calibrating parameters is usually a continuous problem while the ranking of forecasting models is a discrete one. We refer to Gneiting and Raftery (2007) and Gebetsberger et al. (2018) for a description of scoring rules as estimation methods.

method based on the cost of false positives. This is further studied by Merkle and Steyvers (2013) who use the beta family to compare forecasts on various other binary problems. It remains unclear if this approach can be generalized to a non-binary setting. Similarly, Johnstone et al. (2011) propose a class of proper scoring rules that are adapted to the utility function of a decision maker. This yields scoring rules which are restricted to simple settings and may not be expressible in analytical form.

Selection of multivariate proper scoring rules, which we discuss further in Chapter 7, has mostly been limited to simple simulations settings. Pinson and Tastu (2013) evaluate the discrimination ability of the energy score but restrict themselves to a bivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = (\mu, \mu), \qquad \boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

as the data generating process (DGP). Ensembles of 1,000 samples are generated from various misspecified Gaussian distributions and evaluated using 1,000 realisations from the true distribution. Erroneous forecasts differ either in mean, variance or correlation from the ideal one and vary around the correct parameters. The attractiveness of a multivariate scoring rule is then quantified through a discrimination heuristic which measures the average relative distance of the sub-optimal scores from the score obtained by the true distribution. From the magnitude of the relative changes, the authors conclude that the energy score is able to discriminate errors in mean well but lacks sensitivity to errors in variance and especially errors in correlation. Even in the worst case considered, where a perfect correlation is mistaken for zero-correlation, the energy score changes only by 7%.

Similarly, Scheuerer and Hamill (2015) compare their variogram score for $p = 0.5, 1, 2$ with the Dawid-Sebastiani and the energy score in 5 or 15 dimensions, but consider only a DGP with a Gaussian distribution or a Poisson distribution. They compare forecasts that are misspecified in either mean, variance, correlation. A simulation study with ensembles of size 20 or 100 coupled with 5,000 observations from the DGP assesses the ability of each score to identify the correct model through

the score sample mean. To test the robustness, each simulation is repeated 10 times. The authors then quantify the discrimination ability by examining the rankings. A good scoring rule should be able to identify the correct DGP every time whereas a bad one chooses a misspecified model in at least one run. Overall, Scheuerer and Hamill (2015) confirm the finding of Pinson and Tastu (2013) that the energy score lacks sensitivity to misspecification in the dependency structure. Both variogram score with $p = 0.5$ and $p = 1$ perform well and are able to identify the true DGP while the performance of the Dawid-Sebastiani score is mixed because the ensembles are too small for accurate covariance estimations. This comparison of multivariate scoring rules is more comprehensive than alternative ones, but still mainly uses Gaussian distributions as the DGP. In the case where a Poisson distribution is assumed as the DGP, all scores but the variogram score with $p = 0.5$ had at least some ranking issues and may identify the wrong model as the correct one.

Most research in the forecasting literature in finance and economics includes a short empirical study, but extensive application of proper scoring rules, even to univariate distribution forecasts of financial returns, are hard to find. This is possibly due to their computational complexity. Many papers in the GARCH forecasting literature base out-of-sample tests only on point forecasts associated with specific quantiles, especially Value-at-Risk (VaR).[3] Although Zhang and Nadarajah (2018) summarizes a plethora of other point forecasting evaluation methods which have been applied to VaR, the conditional coverage tests introduced by Christoffersen (1998) are by far the most common. These tests are applied by Haugom et al. (2016) to predict VaR using daily, weekly and monthly historical volatility; by Steen et al. (2015) who compare the accuracy of commodity VaR based on predictions from various models; by Bunn et al. (2016) to evaluate forecasting accuracy of extreme quantiles for spot electricity price distributions; and by Clements et al. (2008), who also use the Diebold and Mariano (1995) tests to assess whether different sets of VaR forecasts for exchange rates differ significantly. These tests are also used, in

[3]The voluminous literature on multivariate GARCH forecasting in financial markets is summarised in the useful reviews in Andersen et al. (2006), and Zakamulin (2015).

the context of forecasting median tail loss for setting margins on futures contracts, by Alexander et al. (2019) who also apply the weighted CRPS and the MCS to 16 different univariate GARCH and moving average models, and 12 different multivariate models, using daily returns on the term structure of crude oil, S&P500, gold and Euro/USD exchange rate futures.

The picture is even more incomplete when we consider multivariate distribution forecasting. Much of the large literature on multivariate GARCH models only considers in-sample specification tests – see Silvennoinen and Terasvirta (2009) for a review. An exception is Laurent et al. (2012), who apply MCS and SPA tests to multivariate GARCH forecasts of US stocks based on four different loss differentials between the realised covariance and the model covariance. It is notable that their results are driven by short periods of high market instability during which multivariate GARCH models appear to be inaccurate. Also, they are assessing models by examining the accuracy of covariance matrices, not by utilizing proper multivariate scoring rules to returns themselves. Hence, although these methods have been applied to atmospheric modelling, and to ensemble weather forecasts in particular (Keune et al., 2014) we have not found any previous research which applies them to forecasting multivariate distributions for financial asset returns.

2.2 Distribution Forecasting with Quantiles

Several papers in the distribution forecasting literature apply the quantile regression model of Koenker and Bassett Jr (1978) in a similar way to our forecasting methodology in Chapter 4, but most focus only on certain quantiles. These studies explore univariate distributional effects, but they do not forecast the entire distribution and their evaluation methodology is restricted to single-quantile predictions. In contrast, studies with a focus on distribution forecasts derive the shape of the entire future conditional distribution function by estimating a sufficiently dense grid of quantiles. For the remainder of this chapter, we discuss relevant papers that use quantile regression in a distribution forecasting context and denote $\hat{Q}_\tau(y_{t+1}|\mathcal{I}_t)$ as the τ-quantile estimate of y_{t+1}, conditional on some information $\mathcal{I}_t$ up to time t – see Section 3.1 for a more detailed discussion.

We segment the literature depending on the type of model that is suggested. In Section 2.2.1, we discuss lagged models that condition the quantile forecasts on past information. These methods are easy to use since they do not rely on externally generated forecasts, but generally have a weaker fit compared to contemporaneous alternatives. Cenesizoglu and Timmermann (2008), Zhu (2013) and Pedersen (2015) apply forecast averaging with simple, one-factor models while authors such as Hua and Manzan (2013), Manzan (2015) and Meligkotsidou et al. (2019) consider lagged conditional quantile models with multiple factors.

Contemporaneous models in Section 2.2.2 rely on external sources for the prediction of the explanatory variables which raises the difficulty to utilize them for forecasting. Accurate predictions of the conditional quantiles are dependent on the availability and quality of the external forecast. Gaglianone and Lima (2012) and Bunn et al. (2016) use forecasts of the predictors from various sources. In contrast, Taylor (1999) applies GARCH models to predict the explanatory factors and Ma and Pohlman (2008) circumvent the problem entirely by relying on strong assumptions for the forecast of the conditional quantiles.

In Section 2.2.3, we review some related strands of the forecasting literature. Koenker and Bassett (2010) and Taylor (2007) use quantile regression models without predictor variables to forecast the distribution of scores within basketball games and supermarket sales respectively. Further, Koenker and Leorato (2015) compare the distribution estimation method with conditional quantiles against alternative methods that estimate the distribution directly. We also briefly discuss other empirical studies on exchange rates, interest rates or commodities, but these are restricted to point forecasts.

2.2.1 Lagged Forecasting Models

Cenesizoglu and Timmermann (2008) suggest two lagged quantile regression models for monthly S&P500 returns conditional on $\mathcal{I}_t$. The first model uses only one predictor variable x_t to describe the returns y_t through

$$\hat{Q}_\tau(y_{t+1}|x_t) = \beta_0 + \beta_1 x_t, \tag{2.1}$$

whereas the second one additionally includes an autoregressive term as well as the absolute value of last period's return as common factors. This leads to the dynamic relationship

$$\hat{Q}_\tau(y_{t+1}|\mathcal{I}_t) = \beta_0 + \beta_1 x_t + \beta_2 \hat{Q}_\tau(y_t|\mathcal{I}_{t-1}) + \beta_3 |y_t|. \tag{2.2}$$

A total of 16 variables are considered as predictors, each yielding a separate quantile forecast with Equation 2.1 and 2.2.[4] To augment the simple regression model in Equation 2.1 with multivariate information, the quantile forecasts for each of the predictors are combined with equal weighting. Distribution forecasts are generated through a simple step-function as suggested by Koenker and Bassett (1982) with eleven conditional quantile forecasts spread over the interval $(0, 1)$. The authors then compare the forecasting performance out-of-sample between (i) an equally weighted combination of quantiles from Equation 2.1 across all predictors, (ii) dynamic quantiles of Equation 2.2 based on one of the predictor variables each, (iii) a GARCH(1,1) with normal innovations and (iv) a prevailing quantile model that takes the same form as Equation 2.1 but includes no predictor. Besides an assessment of coverage probabilities and the average loss under the tick loss function, Cenesizoglu and Timmermann (2008) also consider an operational test in portfolio selection where an investor with a power utility allocates wealth to a stock or a risk-free asset. Furthermore, they derive option trading strategies based on the quantile forecasts. The two specifications of the quantile models perform well in

[4]The same common factors have previously been used by Goyal and Welch (2003) to evaluate the predictability of the equity premium.

both statistical and operational tests where they are superior to the GARCH(1,1) and the prevailing quantile benchmark.

Following the approach by Cenesizoglu and Timmermann (2008), Zhu (2013) models monthly returns of the Russel 1000 index and US aggregate bond index. He forecasts quantiles through Equation 2.1 and uses eleven factors for the stock returns and six factors for the bond returns. Again, the quantile forecasts are averaged with equal weighting in both asset classes. Similarly to our Factor Quantile approach, the marginals distributions are then transformed to a joint distribution through a Gumbel copula. A portfolio optimization example based on the Omega ratio suggests an application for the bivariate distribution but Zhu (2013) evaluates neither the performance of the portfolio nor the statistical accuracy of the distribution and leaves that to future research.

Pedersen (2015) applies the techniques proposed by Chakraborty (2003) in combination with the simple factor model in Equation 2.1 to examine a bivariate distribution for the returns of the S&P500 index and a mid-range maturity government bond. Each quantile τ is modelled through the lagged linear regression model where the explanatory factor x_t is one of eight commonly used predictor variables. The grid of quantiles based on one specific predictor is subsequently combined with a step-function to construct a marginal distribution function. This yields eight marginal distributions for both stock and bond returns. The author then uses the coverage tests of Christoffersen (1998) on ten evenly spaced intervals over $(0, 1)$ as well as the weighted logarithmic scoring rule by Amisano and Giacomini (2007) to compare the performance of each marginal distribution forecast against a normal distribution. Additionally, a multivariate analysis is conducted on five quantile combinations to approximate the joint distribution but no forecasting evaluation is provided. It is important to point out that this approach cannot generate joint distribution forecasts. Despite the claims by Chakraborty (2003), that "the geometric quantile process uniquely determines the population distribution, just like univariate quantiles in the

univariate situation", this is generally not the case.[5] Therefore, the multivariate approach does not approximate the joint distribution function which also explains the large difference between univariate and multivariate results in the empirical study.

Lagged quantile models with multiple predictors without forecast averaging are used, for instance, by Hua and Manzan (2013) to forecast the quantiles of high-frequency returns. This leads to

$$\hat{Q}_\tau(y_{t+h}|\mathbf{x}_t) = \beta_0 + \beta_1 \mathbf{x}_t, \tag{2.3}$$

where $\mathbf{x}_t$ contains several volatility measures. Ten versions using different lagged predictors are compared against various asymmetric GARCH models with normal, Student-t and empirically distributed innovations on the S&P500 index and 30-year US treasury bond futures for $h = 1, 2, 5$. Contrary to most other studies, they employ a long out-of-sample period containing 2,419 observations. Hua and Manzan (2013) then construct distributions through kernel density estimation and measure the performance of the predictions with the CRPS and the logarithmic score through the Amisano and Giacomini (2007) test. The quantile model performs well for the stock returns where several versions outperform the GARCH models for $h = 1$. However, none of the quantile model specifications manages to beat the benchmark for the treasury bond future returns and the relative accuracy reduces drastically with the forecast horizon h.

Manzan (2015) uses a quantile autoregressive (QAR) model by Koenker and Xiao (2006) with a panel of 143 lagged variables to forecast the h-month percentage change of four macroeconomic indices. In total, he considers three different versions of an augmented QAR model

$$\hat{Q}_\tau\left(y_{t+h}|\mathbf{x}_t\right) = \alpha(\tau) + \sum_{i=1}^{p_\tau} \beta_i(\tau) y_{t-i+1} + \sum_{j=1}^{J} \gamma_k(\tau) x_{k_j t}, \tag{2.4}$$

[5] Koltchinskii (1997) establishes conditions under which the geometric quantiles characterise the joint distribution.

where the number of lags p_τ is determined by a Schwarz-like criterion and $x_{k_1 t}, \ldots, x_{k_J t}$ are a subset of the panel of predictors $\mathbf{x}_t$.

(i) The first version uses the approach proposed by Stock and Watson (2002) to include the first few principal components. These are derived from the covariance matrix of the macroeconomic predictors.

(ii) A second version chooses variables from the panel of predictors with the LASSO algorithm by Tibshirani (1996). This algorithm is adapted by Koenker (2004, 2011) for quantile regressions. The LASSO penalty is chosen with an approach suggested by Belloni and Chernozhukov (2011).[6]

(iii) The third version combines LASSO and PCA. It takes the relevant explanatory variables selected by LASSO and then further reduces the number of factors in the regression through principal component analysis. Of course, this is only helpful in case the number of predictors selected by LASSO remains large.

Several versions of all models are compared against an autoregressive model with stochastic volatility through proper quantile scores for $h = 3, 6, 12$. The results indicate that the panel data improves the forecast accuracy, especially when LASSO is applied. However, significant outperformances at 5% against the benchmark model are infrequent and only happen in 25% of cases for 3-month ahead forecasts, 39% for 6-month ahead forecasts and 55% for 12-month ahead forecasts, compared to 16%, 27% and 50% for a simple QAR model without any factors.

The lagged augmented QAR model in Equation 2.4 is also studied by Meligkotsidou et al. (2019) to analyse the distribution of US stock market volatility. Instead of LASSO, they combine quantile forecasts with the complete subset regressions by Elliott et al. (2013) and an equally weighted or a stochastic Bayesian combination scheme. The accuracy of this QAR model based on a set of 13 macroeconomic and financial predictors is evaluated on 948 monthly observations against a normal

[6]Quantile regressions with predictors selected by LASSO are also studied by Lima and Meng (2017). They use the lagged, multivariate quantile model in Equation 2.3 but do not explore the distributional effects. Instead, quantiles forecasts are combined to a point forecast similar to Ma and Pohlman (2008).

distribution and a simple QAR model without any factors. Overall, Diebold and Mariano (1995) tests using quantile and logarithmic scores assign superior predictive abilities to the QAR model with forecast combinations. However, the results of the logarithmic score are not significant for the Bayesian combination scheme and only significant at 5% – but not at 1% – for the equally weighted combination scheme, despite using a benchmark model that is encompassed by the augmented QAR model.

A related concept to the LASSO approach is quantile boosting by Fenske et al. (2011) which uses the gradient boosting algorithm by Friedman (2001) in a quantile regression context to select explanatory variables. This has been applied by Pierdzioch et al. (2016) who forecasts monthly gold returns using up to ten lagged predictors. Their empirical study contains data from 1987–2014 and calibrates the model on a rolling window with 60 or 120 observations. The model is evaluated operationally based on the performance of trading strategies against a buy-and-hold investor and yields similar results to LASSO.

2.2.2 Contemporaneous Forecasting Models

Gaglianone and Lima (2012) apply quantile regression to predict the h-step ahead distribution of quarterly U.S. unemployment rates. Their contemporaneous single-factor model

$$\hat{Q}_\tau \left(y_{t+h} | c_{t+h} \right) = \alpha_0(\tau) - \alpha_1(\tau) c_{t+h}$$

generalizes the point forecast model by Capistrán and Timmermann (2009) and estimates the distribution of the unemployment rates through an exogenous consensus forecast c_t^h. This consensus is derived as the average of several point forecasts on the h-step ahead expectation of unemployment rates published by the Survey of Professional Forecasters (SPF). A density function is constructed by fitting an Epanechnikov kernel to the conditional quantile forecasts and examined through 167 observations covering the period from Q1 1969 to Q3 2010. For the evaluation, the authors apply two tests to show the validity of their methodology with $h = 1, 2, 3, 4$:

(i) A Kolmogorov test, adapted by Koenker and Xiao (2002) for robust inference on quantile regression models, does not reject the model specification. Using this result, Gaglianone and Lima (2012) argue that their model is an accurate approximation of the true probability. However, not rejecting the null is of course no indicator of accepting it and may very well be just due to the small sample size which contains only 167 observations.

(ii) A total of 77 out-of-sample forecasts of the quantile model are generated through calibration with an expanding window and then evaluated with the Gaglianone et al. (2011) test. This test focuses on the accuracy of certain quantiles similarly to the coverage test by Christoffersen (1998) with a null hypothesis of accurate quantile forecasts. Again, the test does not reject the null for the quantile model. A symmetric GARCH(1,1) with Gaussian innovations, in contrast, fails in 39% of the cases.

A quantile regression model analytically motivated by a Taylor expansion on the variance forecast of GARCH models is considered by Taylor (1999). He forecasts the quantiles of three exchange rates against the USD over the next h-periods as

$$\hat{Q}_\tau(y_{t+h}|\hat{\sigma}_{t+1}) = \alpha_i(\tau) + \beta_i(\tau)h + \gamma_i(\tau)h\hat{\sigma}_{t+1} + \delta_i(\tau)h^{1/2}\hat{\sigma}_{t+1},$$

where the volatility forecast $\hat{\sigma}_{t+1}$ is obtained using a GARCH(1,1) with Gaussian innovations. Taylor (1999) then compares the accuracy of his model against exponential smoothing and GARCH models on a four year window from 1990 – 1994 with 500 out-of-sample observations where h ranges from 1 to 15 days. Forecasting accuracy is measured by the coverage for the 0.95- and 0.99-quantile estimates and indicates a comparable performance between the quantile model and the benchmarks.

In a rare application of contemporaneous multivariate regression, Ma and Pohlman (2008) combine several common factors to generate point forecasts of stock returns. They start with a contemporaneous linear quantile regression model

$$\hat{Q}_\tau\left(y_{t+1}|\mathbf{x}_{t+1}\right) = \mathbf{x}'_{t+1}\boldsymbol{\beta}(\tau),$$

where $\mathbf{x}_t$ consists of ten variables commonly used by Fama and French (1993) and similar studies to describe stock returns. Although the inclusion of several explanatory variables improves the accuracy of the linear representation, prediction with this model now requires forecasts of the explanatory variables to consider their underlying dependency structure. This increases the difficulty to use the model directly for forecasting greatly, especially if the joint distribution of the common factors is non-elliptical in which case the dependency requirements go beyond the correlation. To circumvent this, Ma and Pohlman (2008) introduce two specifications of their model:

(i) The first version assumes that the conditional location of the stock returns does not change. Therefore, if the realisation at time t is in the lower tail of the distribution, then the prediction for $t+1$ draws a value around the same position of the distribution. Of course, this assumes a strong momentum

effect that may not be realistic in most applications and produces a quantile as point-forecast rather than the mean or median.

(ii) Alternatively, they set the point forecast as a weighted sum of the quantiles. Given a quantile partition $\mathbb{Q} = (\tau_1, \ldots, \tau_n)$ with ascending values for which the quantiles are estimated, this yields the prediction

$$\hat{y}_{t+1} = \sum_{i=1}^{n} (\tau_i - \tau_{i-1}) \hat{Q}_{\tau_i} \left(y_{t+1} | \mathbf{x}_{t+1} \right), \qquad \tau_0 = 0. \tag{2.5}$$

However, assuming the quantile model is a good representation of the actual quantiles, this just estimates the future stock return through a discrete approximation of

$$\int_0^1 \hat{Q}_\tau \left(y_t | \mathbf{x}_t \right) \, \mathrm{d}\tau,$$

which is the expectation of the stock return at time t. The models are not evaluated statistically nor operationally. Instead, the authors provide a theoretical result to show that the mean absolute deviation (MAD) of their forecasting methodology is not higher than that of alternative, traditional estimations for mean or median.

Combining lagged and contemporaneous explanatory variables, Bunn et al. (2016) use the general quantile regression by Chernozhukov and Umantsev (2001) to model future conditional quantiles of the UK electricity price between 18:30 – 19:00 as

$$\hat{Q}_\tau \left(y_{t+1} | \mathbf{x}_t, \mathbf{x}_{t+1} \right) = \mathbf{x}_t' \boldsymbol{\beta}_1(\tau) + \mathbf{x}_{t+1}' \boldsymbol{\beta}_2(\tau).$$

Forecasts for the non-lagged predictors are provided by the System Operator, who presumably considers the dependency structure between the forecasted variables.[7] The unconditional coverage test by Kupiec (1995) and the conditional coverage test by Christoffersen (1998) are used to evaluate the models in a 5-year window

[7] The quantile model utilizes the demand and reserve margin forecast but it is unclear how the System Operator produces those predictions.

with 1,185 out-of-sample observations but focusing only on the left and right tail of the distribution. The authors compare several specifications of their model, that differ with respect to the choice of predictors, against a GARCH(1,1) with normal or t-distributed innovations, as well as a conditional autoregressive Value-at-Risk (CAViaR) model by Engle and Manganelli (2004). In this evaluation, quantile models provide slightly superior accuracy relative to the benchmark models.

2.2.3 Alternative Forecasting Models

Koenker and Bassett (2010) apply a parsimonious quantile regression model in a distribution forecasting setting to predict the scores of basketball games during the NCAA Division I Men's Basketball Tournament (March Madness). Using all information up to time t, the distribution for the final score of team i against team j in a game is approximated as

$$\hat{Q}_\tau(y_{ij}|\mathcal{I}_t) = \alpha_i(\tau) - \delta_j(\tau)$$

with constant offensive rating α_i of team i and constant defensive rating δ_j of team j. Contrary to most other model specifications in the literature, no predictor variables are included and all factors in the regression model are assumed to be constant. This specification improves upon the point forecast of the conditional score expectation

$$\mathbb{E}(y_{ij}|\mathcal{I}_t) = \alpha_i - \delta_j$$

by considering the entire distribution and allowing for asymmetric effects. It is also less sensitive to outliers and requires fewer assumptions.[8] The authors combine a dense grid of 199 equally-spaced quantiles in $(0, 1)$ through kernel density estimation with a Gaussian kernel into marginal distributions for the scores of team i and j. A Frank copula accounts for possible dependency between the scores within one game. This model is estimated on 2,940 games for each of the 232 Division I NCAA teams from 2004 – 2005. This requires the estimation of 464 parameters for each of the 199 quantiles which is only possible for such a simple model. The resulting distributions are then evaluated operationally with 48 out-of-sample games by betting whether the realized point spread or point sum is higher or lower than the ex-ante announced estimates by the bookie. A bet on a combination of both point spread and point sum is also considered. The quantile model achieves mildly favourable performance, beating the bookie in 57% of the cases for the point spread and the point sum as well as in 27% for the parley combination. These odds are an improvement in

[8] Particularly the independency assumption of ordinary least squares is questionable in the data set because of momentum effects.

comparison to random guessing which yields 50% frequency of success for the point spread or point sum and 25% for the two-stage bet. However, despite accounting for heterogeneous effects, the quantile model fails to outperform the simpler least squares alternative.

In a related strand of literature, Taylor (2008) generalizes exponentially weighted least squares (EWLS) to a quantile regression context. The exponentially weighted quantile regression (EWQR) adapts the minimization problem for ordinary quantile regression with a decay parameter and can be formulated as a linear program. He shows that the resulting EWQR estimator $\hat{Q}_\tau\left(y_t|\mathcal{I}_t\right)$ can estimate the conditional distribution of y_t through

$$\hat{F}\left(\hat{Q}_\tau\left(y_{t+1}|\mathcal{I}_t\right)\right) = \frac{\sum_{i=1}^{t+1}\lambda^{t+1-i}\mathbb{1}\left\{y_i < \hat{Q}_\tau\left(y_i|\mathcal{I}_i\right)\right\}}{\sum_{i=1}^{t+1}\lambda^{t+1-i}},$$

where $\lambda \in [0,1]$ is a weighting parameter. The model can be extended by an additional term or dummy variable to incorporate trends or seasonality. Taylor (2007) uses the distribution from EWQR quantiles to forecast supermarket sales but only considers point forecasts.

The general ability of quantile regression to estimate accurate distribution functions is examined by Koenker and Leorato (2015). They compare distribution estimations of a random variable Y, either indirectly through the conditional quantiles or directly through estimations of the conditional mean of binary indicators as

$$D_i = \mathbb{1}\{Y \leq y_i\} \tag{2.6}$$

at a finite number of cut-off values $y_1, \ldots, y_n$. The direct approach is initially suggested by Foresi and Peracchi (1995) and is applied by Rothe (2012) and Chernozhukov et al. (2013), amongst others. Both approaches estimate the distribution but differ in their methodologies. Koenker and Leorato (2015) analyse the resulting distributions asymptotically and through Monte Carlo experiments, assuming either correctly specified models, or misspecified ones with small or large regression

R^2. They show that, under general assumptions, both methods are asymptotically equivalent in terms of their asymptotic relative efficiency (ARE) if the models are correctly specified. However, in a simulation setting, the finite sample performance favours the quantile regression estimation. Both the direct and indirect approach are similar with small regression R^2 but as the R^2 increases, the indirect method with quantile regression is more efficient than the direct approach.

Most other recent empirical studies of forecasting in exchange rates, interest rates or commodities only consider point forecasts. Particularly important to our study are those using latent factors, such as Greenaway-McGrevy et al. (2018) who use principal component analysis (PCA) to identify only two latent factors driving US exchange rates. They only evaluate out-of-sample point forecasts compared with the random walk benchmark. For commodities the two most relevant papers are by Zolotko and Okhrin (2014) and Chen et al. (2014). Zolotko and Okhrin (2014) model the joint time-series dynamics of natural gas and heating oil forward curves, whereas we use a broader cross section of Bloomberg investable indices. Their focus is on risk management, so they quantify forecasting accuracy using portfolio Value-at-Risk, not the entire distribution. Chen et al. (2014) analyse common components in a large panel of relative commodity prices, comparing out-of-sample point forecasts using the Diebold and Mariano (1995) test in addition to root mean square errors (MSE). Finally, for forecasting the US treasury yield curve, Almeida et al. (2017) also advocate the use of latent factors. They evaluate forecasting accuracy of models based on a segmented error-correction framework, relative to random walk and autoregressive alternatives, using methodologies similar to Chen et al. (2014) – but again they only consider point forecasts.

Chapter Three

Theoretical Background

3.1 Quantile Regression 33
3.2 Principal Component Analysis 38
3.3 Copulas 42
3.4 Univariate GARCH Models 48
3.5 Multivariate GARCH Models 51

In this chapter we introduce the necessary theoretical background for our semi-parametric multivariate Factor Quantile model and our benchmarks in Chapters 6 and 7.

We start with the theory of quantile regression in Section 3.1 which we use to construct the marginal Factor Quantile distributions. A brief summary of definitions and associated quantile regression methodologies points out the issues associated with quantiles in higher dimensions.

Then, we describe principal component analysis in Section 3.2, with a derivation of the principal component representation that we apply as statistical factor model. These versions of Factor Quantile models are especially useful if a fundamental or macroeconomic model is not available or if the common factors are difficult to forecast.

Section 3.3 defines general dependency measures based on concordance and covers popular copula parameterisations which extend our non-parametric marginal distributions to a joint distribution function.

Last, Sections 3.4 and 3.5 provide the theory for univariate and multivariate GARCH models that we use in our empirical and simulation study in Chapters 6 and 7 as benchmarks. We define constant and dynamic conditional correlation GARCH models that apply E-GARCH(1,1) with Student-t distributed innovations as their univariate basis. A discussion on complexity justifies our choices for the multivariate GARCH specifications in the subsequent chapters.

3.1 Quantile Regression

Quantile regression was developed by Koenker and Bassett Jr (1978) to estimate conditional quantiles through optimization. In contrast to ordinary least squares, effects outside the mean can be captured. The method is also more robust towards outliers and can be implemented efficiently using linear programming and the simplex algorithm (Koenker and d'Orey, 1987).

Definition 3.1 (Quantiles)**.** Let Y be a random variable with distribution F. Then, for any $0 < \tau < 1$,

$$F^{-1}(\tau) = \inf \{y : F(y) \geq \tau\}$$

is called the τ-th quantile of Y.

Contrary to ordinary least squares regression for which the factor loadings have analytical solutions under general assumptions, quantile regression estimates arise from optimization with the asymmetric penalty function

$$\rho_\tau(u) \coloneqq u(\tau - \mathbb{1}\{u < 0\}). \tag{3.1}$$

We seek $\hat{y}$ that minimizes expected loss

$$\mathbb{E}\left(\rho_\tau(Y - \hat{y})\right) = (\tau - 1) \int_{-\infty}^{\hat{y}} (y - \hat{y}) \, \mathrm{d}F(y) + \tau \int_{\hat{y}}^{\infty} (y - \hat{y}) \, \mathrm{d}F(y).$$

This requires the first-order condition

$$0 \stackrel{!}{=} (1 - \tau) \int_{-\infty}^{\hat{y}} \mathrm{d}F(y) - \tau \int_{\hat{y}}^{\infty} \mathrm{d}F(y) = F(\hat{y}) - \tau$$

for minimization. Therefore any element in $\{y : F(y) = \tau\}$ is a solution which in turn means that the minimization problem derives the τ quantile of Y. If the set contains more than one element, we choose the smallest one to ensure a left-continuous quantile function.

In a linear factor model context, the conditional τ-th quantile of the dependent variable Y are described through random variables $\mathbf{X} = (X_1, \ldots, X_m)'$ as

$$Q_Y(\tau|\mathbf{X}) = \alpha + \boldsymbol{\beta}\mathbf{X} + \varepsilon, \tag{3.2}$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$. The factor loadings associated with the conditional τ-th quantile can hence be estimated through

$$\left(\hat{\alpha}, \hat{\boldsymbol{\beta}}\right) = \underset{(\alpha,\boldsymbol{\beta})}{\arg\min}\, \mathbb{E}\left(\rho_\tau\left(Y - \alpha - \boldsymbol{\beta}\mathbf{X}\right)\right)$$

with the associated residuals

$$\hat{e} = Y - \hat{\alpha} - \hat{\boldsymbol{\beta}}\mathbf{X}.$$

Consequently, the conditional quantiles are calculated by optimization rather than sorting. Replacing the expectation with the sampling mean yields the sample quantiles loadings

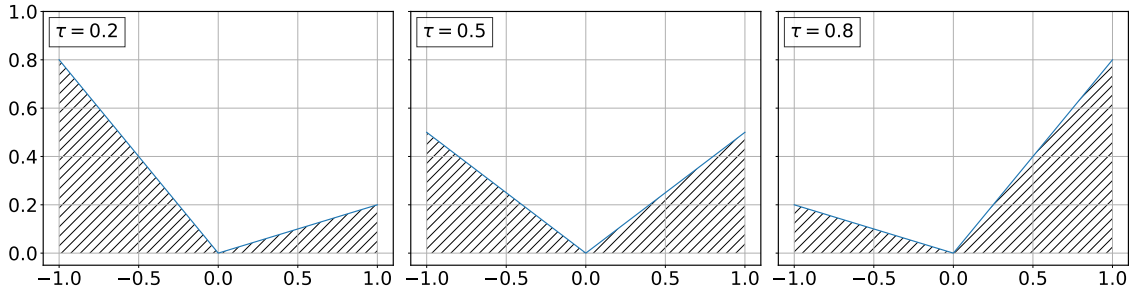
$$\left(\hat{a}, \hat{\boldsymbol{b}}\right) = \underset{(a,\mathbf{b})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(y_i - \hat{a} - \hat{\boldsymbol{b}}\mathbf{x}_i\right),$$

where $(y_i)_{i=1}^n$ and $(\mathbf{x}_i)_{i=1}^n = (x_{i1}, \ldots, x_{im})_{i=1}^n$ are observations of the dependent and the independent variables. Koenker and Bassett Jr (1978) prove that the sample quantiles corresponding to those factor loadings are asymptotically consistent and jointly normally distributed given some regularity assumptions.[1] Furthermore, Angrist et al. (2006) show that quantile regression minimizes the expected weighted mean-squared error and therefore is the best linear approximation to the unknown conditional quantile function whereas ordinary least squares best approximates the conditional expectation function.

Quantile regression gives asymmetric penalties for under-prediction and over-prediction as depicted in Figure 3.1. This loss function results in much more robust estimates where the factor loadings of the regression are less influenced by outliers in the data. In fact, given any non-negative $c \in \mathbb{R}$ and the quantile regression model in Equation 3.2, then

[1] We refer to Koenker (2005, pg. 116–124) for a summary of the assumptions. Notably, these results also hold even if the errors are not iid.

Figure 3.1: Asymmetric penalty for quantile regression with various parameters

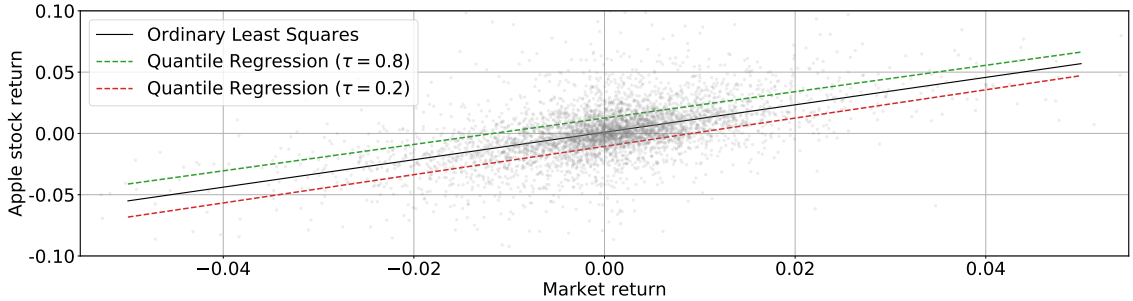


The asymmetric response function in Equation 3.1 is depicted with various parameters. Negative deviations between realisation and prediction are penalized with a slope of $\tau - 1$ while positive deviations are penalized with a slope of τ.

$$\left(\hat{\alpha}, \hat{\boldsymbol{\beta}}\right)(\tau, Y, \mathbf{X}) = \left(\hat{\alpha}, \hat{\boldsymbol{\beta}}\right)(\tau, \alpha + \boldsymbol{\beta}\mathbf{X} + c\hat{e}, \mathbf{X}),$$

where $(\hat{\alpha}, \hat{\boldsymbol{\beta}})(\tau, Y, \mathbf{X})$ are the factor loadings belonging to quantile τ for a quantile regression between Y and $\mathbf{X}$. This means quantile regression yields the same factor loadings for any regressand as long as the sign of the residual stays the same. Figure 3.2 illustrates this perhaps unintuitive result geometrically in a sample with low dimensions. Given some data cloud, quantile regression fits a hyperplane that segments the data into two sub-spaces, containing approximately τ and $1 - \tau$ percent of the data respectively. Changing any observation in the data cloud without crossing the hyperplane does not impact the relative split of the data and hence induces no need to adjust the hyperplane.

Figure 3.2: Quantile regression hyperplane for CAPM on Apple stock returns



We use the Capital Asset Pricing Model (CAPM) to explain Apple stock returns. The market return is approximated by the return of the S&P500 index. Factor loadings of the ordinary least squares and the quantile regressions are based on daily data ranging from 01 January 2000 to 31 December 2018. Approximately 20% of the data is above the green line and 80% of the data is above the red line.

A problem may arise with quantile regressions on different τ as a direct result of the robustness. In case the quantiles are estimated independently, there is no guarantee that the conditional quantiles $Q_Y(\tau|\mathbf{X})$ are monotonically increasing with τ because only information in a neighbourhood of the separating hyperplane are crucial for the calibration. Solely in the special case where the quantiles are conditioned on the mean of the explanatory variables $\overline{\mathbf{X}}$, the resulting sample quantiles $\hat{Q}_Y(\tau|\overline{\mathbf{X}})$ are always non-decreasing in τ (Koenker, 2005). Violations of monotonicity are referred to as the quantile crossing problem and typically only occur in the outer regions of the design space (Koenker and Bassett, 1982). Several solutions have been proposed including the location-scale shift model by He (1997) and monotone rearranging by Chernozhukov et al. (2010).

Extending quantile regression to higher dimensions presents a challenge. The basic problem is that the definition of a multivariate quantile is not unique as there is no inherent ordering in multiple dimensions. Various approaches exist:

(i) Following Tukey (1974), Chaudhuri (1996) introduces the notation of geometric multivariate quantiles, in terms of their distance and direction from the centre of the data cloud of observations on the variables, using elements in the open unit ball to extend the Koenker and Bassett Jr (1978) loss function.

(ii) Alternatively, Chakraborty (2001) regards a quantile vector as that which has marginal quantiles of identical probability to its components, but this ignores any co-dependency between the components in the vector and only applies when the variables are independent.

(iii) Also, Cai (2010) extends the bivariate quantiles of Gilchrist (2000) to particular quantile surfaces for each variable, but leaves the relationship between these to further research.

The varying definitions of multivariate quantiles lead to alternative and conflicting generalizations of quantile regression:

(i) For instance, Chakraborty (2003) proposes to minimize a loss function that is a straightforward multivariate equivalent of the standard loss function used in univariate quantile regression, introduced by Koenker and Bassett Jr (1978). However, this does not allow estimation of an associated distribution function because it is only based on the notion of geometric multivariate quantiles.

(ii) Similarly, Hallin et al. (2010) use the half-space depth contours of Tukey (1974) which are not equivalent to an associated distribution function.

(iii) By contrast, insisting on the equivalence between the quantile function and a well-defined multivariate distribution, Chavas (2018) proposes that a multivariate τ-quantile is a set $\mathbf{c}$ corresponding to the τ-contour of the multivariate distribution F, i.e. $F(\mathbf{c}) = \tau$. This must reflect the general properties of τ-quantiles, e.g. $F(\mathbf{c})$ is always non-decreasing – however, the τ-contours need not be convex and so F need not have a unique inverse. Hence, Chavas (2018) assumes that quantiles are linear functions of exogenous variables. He only derives statistical properties of the quantile estimator when conditional distributions of the endogenous variables are independent.

Quantiles are related to expectiles, a concept introduced by Newey and Powell (1987). Rather than using the $\mathcal{L}^1$-norm in Equation 3.1, expectiles are calculated through the $\mathcal{L}^2$-norm and the asymmetric penalty function

$$\rho_\tau(u) := u^2|\tau - \mathbb{1}\{u < 0\}|.$$

Despite the similar optimization structure, there is no straightforward relationship between quantiles and expectiles of some distribution F outside of simplified settings. In fact, Jones (1994) shows that expectiles of a distribution F themselves are quantiles of a distribution G, which can be expressed through the distribution and partial moments of F. Similarly to quantiles, expectiles describe parts of the underlying distribution with the main difference that expectiles have a global dependence in contrast to the local robustness of quantiles.

3.2 Principal Component Analysis

Principal component analysis (PCA) is an orthonormal linear transformation technique by Pearson (1901) and Hotelling (1933) that turns correlated random variables into a set of orthogonal ones of decreasing variance. As such it has become a popular tool in multivariate data analysis, especially in the presence of high dimensions.

In the literature there are many varying definitions of principal components. While minor deviations arise from the choice of normalization constraints, other more pronounced variations are also in circulation. To build a foundation for further discussion, we define the principal components as follows:

Definition 3.2 (Principal components)**.** Let $\mathbf{x} = (X_1, \ldots, X_d)$ be a random vector. The first principal component of $\mathbf{x}$ is defined as $P_1 := \boldsymbol{\alpha}_1'\mathbf{x}$, where

$$\boldsymbol{\alpha}_1 := \arg\max_{\boldsymbol{\alpha}} \left\{ \mathbb{V}\mathrm{ar}(\boldsymbol{\alpha}'\mathbf{x}) : \boldsymbol{\alpha} \in \mathbb{R}^d, \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1 \right\}.$$

Similarly, the i-th principal component of $\mathbf{x}$ for $2 \leq i \leq d$, is defined as $P_i := \boldsymbol{\alpha}_i'\mathbf{x}$, where

$$\boldsymbol{\alpha}_i := \arg\max_{\boldsymbol{\alpha}} \left\{ \mathbb{V}\mathrm{ar}(\boldsymbol{\alpha}'\mathbf{x}) : \boldsymbol{\alpha} \in \mathbb{R}^d, \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1, \boldsymbol{\alpha}'\mathbf{x} \perp \boldsymbol{\alpha}_j'\mathbf{x} \text{ for all } 1 \leq j < i \right\}.$$

Therefore, the principal components are normed vectors under the $\mathcal{L}^2$-norm which are orthogonal to each other and for which variance decreases with each component. As statistical constructs, principal components do not necessarily have intuitive interpretations. However, since the first component explains most of the variation of $\mathbf{x}$, it is often interpreted as the common trend.

Generally, principal components are not calculated through the optimization in Definition 3.2 but rather by algebraic methods. Let $\mathbf{V}$ be the covariance matrix of $\mathbf{x}$. If its eigenvalues are distinct and non-zero, the i-th principal component of $\mathbf{x}$ is $\boldsymbol{\alpha}_i'\mathbf{x}$ where $\boldsymbol{\alpha}_i$ is the normalized eigenvector under the $\mathcal{L}^2$-norm corresponding to the i-th largest eigenvalue λ_i.[2] Non-distinct or zero eigenvalues are unlikely to occur in a practice but might complicate statistical inference:

[2]We refer to Jolliffe (1986, pg. 5–6) for a proof of this statement.

Some eigenvalues are zero Let k eigenvalues be zero. The rank of $\mathbf{V}$ is $d - k$ and the principal components corresponding to the zero eigenvalue(s) have zero variance. Thus the number of variables can be reduced from d to $(d - k)$ without any loss of information.

Some eigenvalues non-distinct Let k eigenvalues be equal to each other. The corresponding k eigenvectors span a k-dimensional space in which the eigenvectors are arbitrary with the restriction of being orthogonal to one another. This means that the k principal components matching those eigenvectors are not uniquely defined.

The alternative calculation for the principal components $\mathbf{p} = (P_1, \ldots, P_d)$ remains valid when the covariance matrix is replaced by the correlation matrix or any non-equally weighted covariance or correlation matrix. While principal components based on covariance matrices take both the volatilities and the correlation structure of $\mathbf{x}$ into account, principal components based on correlation matrices are only influenced by the correlation. Hence, elements of $\mathbf{x}$ with large variances tend to dominate the covariance matrix based principal components which could distort the results. Since there is no general relationship between the spectral decomposition of the covariance matrix and that of the correlation matrix, there is no general relationship between their respective principal components either. However, in case the volatilities of $X_1, \ldots, X_d$ are similar, the principal components from the covariance matrix and the correlation matrix will also be similar to each other.

We can derive a linear representation of $\mathbf{x}$ through the spectral decomposition for the principal component. Let

$$\mathbf{W} := (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_d')$$

be a matrix of the eigenvectors of $\mathbf{V}$, ordered decreasingly according to their corresponding eigenvalues. Since we assumed the eigenvalues of $\mathbf{V}$ to be distinct, $\mathbf{W}$ is an orthogonal matrix. Therefore, the principal component representation follows from Definition 3.2 with

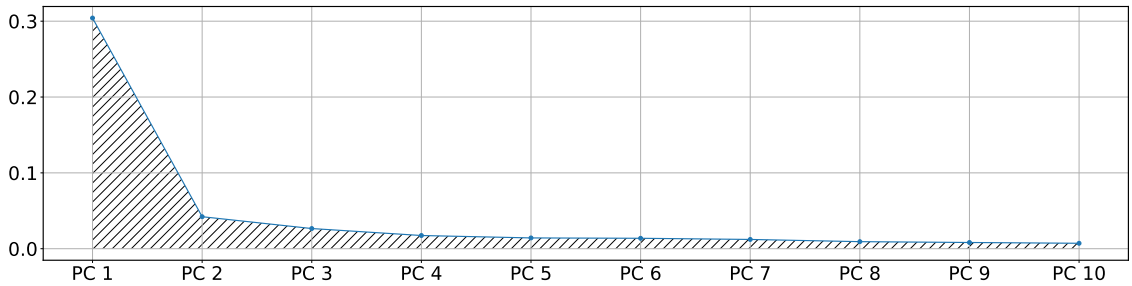
$$\mathbf{x} = \mathbf{W}'\mathbf{p}. \tag{3.3}$$

The principal components each capture a proportion of the variance of $\mathbf{x}$. Let the total variability of $\mathbf{x}$ be defined as $\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(X_i)$. Since $\mathbf{V}$ is a square matrix, it is similar to its Jordan form and the total variability can be expressed as

$$\sum_{i=1}^{d} \mathbb{V}\mathrm{ar}(X_i) = \mathrm{tr}(\mathbf{V}) = \sum_{i=1}^{d} \lambda_i = \sum_{i=1}^{d} \boldsymbol{\alpha}_i'\mathbf{V}\boldsymbol{\alpha}_i = \sum_{i=1}^{d} \mathbb{V}\mathrm{ar}(\boldsymbol{\alpha}_i'\mathbf{x}). \tag{3.4}$$

Thus, the total variance of $\mathbf{x}$ is explained by the principal components in decreasing order and the percentage of the variance of the first $1 \leq k \leq n$ principal components is given by the ratio $(\sum_{i=1}^{k} \lambda_i)/(\sum_{i=1}^{d} \lambda_i)$. The proportion of variance explained by each principal component $\lambda_i/(\sum_{i=1}^{d} \lambda_i)$ is often illustrated as a scree plot as shown in Figure 3.3.

Figure 3.3: Percentage of variance explained by first principal components



The first 10 out of 505 principal components based on the stock returns of the S&P500 constituents are displayed in the scree plot. Those explain 46% of the variability in the data according to Equation 3.4. The covariance matrix is based on daily data ranging from 01 January 2000 to 31 December 2018.

Equation 3.4 is often used to reduce the dimension of the multivariate random vector $\mathbf{x}$. For collinear $\mathbf{x}$, the contribution to the variance explained of the last principal components are minor. Hence, instead of analysing the untransformed, d-dimensional random vector $\mathbf{x}$, we can limit our analysis to the first few components and thereby reduce the dimension from d to some $k < d$ while sacrificing only a relatively low amount of variance explained. It can be shown that the principal component representation with the first k components maximizes the variability

explained while the principal component representation with only the last k factors minimizes the variability explained out of any linear representation with k factors.[3] The choice of k depends on the collinearity of $\mathbf{x}$ and can be determined by either choosing a percentage of the variance explained one wants to capture or graphically through slopes in the scree plot in Figure 3.3.[4]

[3] See Jolliffe (1986, pg. 11–13) for a proof.

[4] A large number of methods have been proposed to determine the optimal number of principal components. We refer to Jolliffe (1986, Chapter 6) for a review.

3.3 Copulas

Copulas are multivariate distribution functions with uniform marginals which can be used to decompose a joint distribution function into its marginals and dependency structure. They are widely applied in finance and economics, especially in problems with higher dimensions (Patton, 2009).

Definition 3.3 (Copula)**.** A d-dimensional copula $C : [0,1]^d \to [0,1]$ is a distribution function of a random vector with uniform marginals $\mathcal{U}(0,1)$.

The main reason for the popularity of copulas is Sklar's theorem. Let F be a multivariate distribution function with marginals $F_1, \ldots, F_d$. Then, there exists a copula $C : [0,1]^d \to [0,1]$ such that

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

C is unique on $\text{ran}(F_1) \times \ldots \times \text{ran}(F_d)$. Conversely, if $C : [0,1]^d \to [0,1]$ is a copula and $F_1, \ldots, F_d$ are univariate distribution functions, then

$$C(F_1(x_1), \ldots, F_d(x_d))$$

defines a joint distribution function with marginals $F_1, \ldots, F_d$.

Sklar's theorem implies that any univariate distributions can be linked with any copula to yield a valid multivariate distribution function. Marginal distributions and dependency structure can therefore be chosen separately and independently. Further, if marginal densities $f_1, \ldots, f_d$ are available, then the multivariate density is given by

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) f_1(x_1) \cdots f_d(x_d), \tag{3.5}$$

where c is the density of the copula.

Patton (2006) extends Sklar's theorem to conditional distributions. Given some information set $\mathcal{I}$, the conditional joint distribution $F(\cdot|\mathcal{I})$ with marginals $F_1(\cdot|\mathcal{I}), \ldots, F_d(\cdot|\mathcal{I})$ can be expressed as

$$F(x_1, \ldots, x_d|\mathcal{I}) = C\left(F_1(x_1|\mathcal{I}), \ldots, F_d(x_d|\mathcal{I})|\mathcal{I}\right)$$

with a unique conditional copula $C(\cdot|\mathcal{I})$. Furthermore, given any conditional marginal distributions $F_1(\cdot|\mathcal{I}), \ldots, F_d(\cdot|\mathcal{I})$ and any conditional copula $C(\cdot|\mathcal{I})$,

$$C\left(F_1(x_1|\mathcal{I}), \ldots, F_d(x_d|\mathcal{I})|\mathcal{I}\right)$$

is a valid conditional joint distribution with conditional marginals $F_1(\cdot|\mathcal{I}), \ldots, F_d(\cdot|\mathcal{I})$. All distributions and copulas must be conditioned on the same information set, otherwise the resulting function $F(x_1, \ldots, x_d|\mathcal{I})$ might not be a well-defined joint distribution (Fermanian and Wegkamp, 2012).

Copulas are generally calibrated through maximum likelihood estimation. In case a non-parametric model is chosen for the marginals and a parametric model for the copula, the estimator is called canonical maximum likelihood. The two-step estimator of Chen and Fan (2006*b*,*a*) should then be used rather than standard estimation methods since the likelihood depends on the marginal non-parametric distributions $F_1, \ldots, F_d$ and their parameters.

The goodness of fit can be estimated by comparing the fitted copula to the empirical copula through the Kolmogorov–Smirnov or the Cramér–von Mises test. Rémillard (2010) shows that the test statistics are unaffected by the estimation of the marginal distributions in the case of non-parametric marginals with parametric copulas. Therefore only estimation errors from the empirical distribution function need to be addressed, for which he proposes a simulation-based method. Alternatively, information criteria such as the Akaike information criterion or the Bayesian information criterion can be applied.

Definition 3.4. Let $\hat{L}$ be the maximized value of the likelihood function with k parameters based on n observations. The Akaike information criterion is defined as

$$\mathrm{AIC} = 2k - 2\log(\hat{L})$$

and the Bayesian information criterion is defined as

$$\mathrm{BIC} = \log(n)k - 2\log(\hat{L}).$$

Dependency is often measured through the Pearson correlation coefficient and much of the applied literature in finance still focuses on this statistic. Embrechts et al. (1999) lists crucial issues with this coefficient which is only a good measure of dependency given elliptical distributions. For other distributions, correlation and marginal distributions alone are not able to determine the joint distribution since there are infinitely many joint distributions that fit the specified criteria. Furthermore, some linear correlations in the interval $[-1, 1]$ can not always be attained. These restrictions motivate the use of other dependency measures in finance and economics, where the distributions are often non-elliptical (Chicheportiche and Bouchaud, 2012). Especially in the context of copulas, the Pearson correlation coefficient is suboptimal since it is also affected by the marginal distributions rather than focusing on the dependency structure imposed by the copula.

Definition 3.5 (Pearson correlation coefficient)**.** Let $(x_i, y_i)_{i=1}^n$ be sample observations of the random variables (X, Y) with sample means $\overline{x}$ and $\overline{y}$. Then, the Pearson correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}}.$$

Spearman's ρ and Kendall's τ are alternative measures of dependency which are based on the concordance. They are functions of the rank of the data only which means only the order during sorting is relevant. Therefore they depend solely on the copula but not the marginal distributions.

Definition 3.6 (Concordance)**.** Let (x_1, y_1) and (x_2, y_2) be two sample observations of the random variables (X, Y). The pair is concordant, if $(x_1 - x_2)(y_1 - y_2) > 0$ and discordant if $(x_1 - x_2)(y_1 - y_2) < 0$.

Definition 3.7 (Spearman's ρ)**.** Let $(x_i, y_i)_{i=1}^n$ be sample observations of the random variables (X, Y) with corresponding ranks $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))_{i=1}^n$. Then, Spearman's ρ is defined as Pearson correlation coefficient between the ranked variables.

Definition 3.8 (Kendall's τ)**.** Let $(x_i, y_i)_{i=1}^n$ be sample observations of the random variables (X, Y). Further, let N_c be the number of concordant pairs in the sample observations and N_d be the number of discordant pairs. Then, Kendall's τ is defined as

$$\tau = \frac{N_c - N_d}{n(n-1)/2}.$$

Another dependency measure that focuses on the concordance in the tails is the tail dependence.

Definition 3.9 (Tail dependence)**.** Let $\mathbf{X} = (X_1, \ldots, X_d)$ with marginal distributions $F_1, \ldots, F_d$. Then, assuming the limits exist, the i, j-th lower tail dependence is defined as

$$\lambda_{ij}^l = \lim_{q \downarrow 0} P\left(X_i < F_i^{-1}(q) | X_j < F_j^{-1}(q)\right)$$

and the i, j-th upper tail dependence as

$$\lambda_{ij}^u = \lim_{q \uparrow 0} P\left(X_i > F_i^{-1}(q) | X_j > F_j^{-1}(q)\right).$$

There is a large literature of copulas. We limit our discussion to relatively simple specifications that are commonly used in practice and refer to Nelsen (2006) for a full review.[5] Figure 3.4 illustrates the densities of the copulas we consider.

One of the simplest copulas is the Gaussian copula. It is symmetric with zero to weak tail dependence unless the correlation is one.

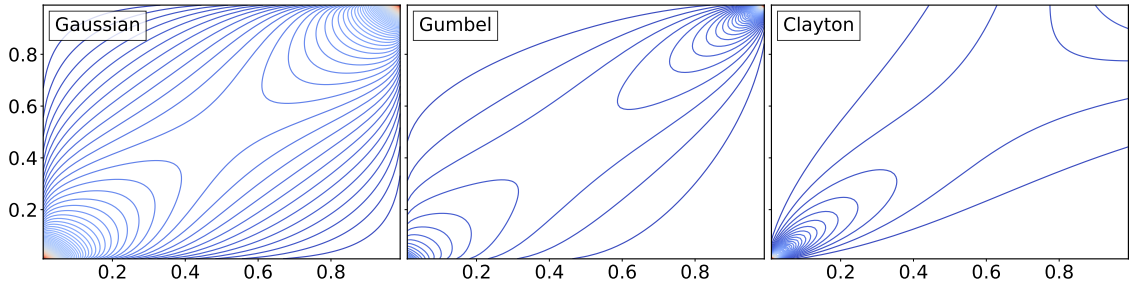
Definition 3.10 (Gaussian copula)**.** Given a correlation matrix $\mathbf{\Sigma}$, the Gaussian copula is defined as

$$C(x_1, \ldots, x_d) = \mathbf{\Phi_\Sigma}(\Phi^{-1}(x_1), \ldots, \Phi^{-1}(x_d)),$$

where $\mathbf{\Phi_\Sigma}$ is a multivariate normal distribution with zero mean and correlation $\mathbf{\Sigma}$. Φ denotes the standard normal distribution.

[5]Notable copulas that work for higher dimensions include nested Archimedean copulas (Hering et al., 2010; Hofert and Scherer, 2011), vine copulas (Aas et al., 2009; Min and Czado, 2010) and factor copulas (Oh and Patton, 2017).

Figure 3.4: Densities of popular copulas ($d = 2$)



The contour plots display the densities of the Gaussian, Gumbel and Clayton copula for $d = 2$. The Gaussian copula has correlation $\rho = 0.6$, while both the Gumbel and the Clayton copula use $\theta = 2$.

The density of a Gaussian copula can be derived through Equation 3.5 as

$$c_{\text{Gauss}}(x_1, \ldots, x_d) = \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(x_1) \\ \vdots \\ \Phi^{-1}(x_d) \end{pmatrix}' \left(\boldsymbol{\Sigma}^{-1} - I \right) \begin{pmatrix} \Phi^{-1}(x_1) \\ \vdots \\ \Phi^{-1}(x_d) \end{pmatrix} \right)$$

but there is no closed form expression for the corresponding distribution function.

Archimedean copulas are a class of copulas which rely only on one parameter θ and are therefore easy to calibrate.

Definition 3.11 (Archimedean copula)**.** A copula C is Archimedean if there exists a generator function $\psi : [0, 1] \to [0, \infty)$ such that

$$C(x_1, \ldots, x_d) = \psi^{-1}(\psi(x_1) + \ldots + \psi(x_d)),$$

where ψ is a d-monotone function with $\psi(1) = 0$ and $\psi(x) \to \infty$ as $x \to 0$.

Densities for Archimedean copulas are given by

$$c(x_1, \ldots, x_d) = \psi_{(d)}^{-1}\left(\psi(x_1) + \ldots + \psi(x_d)\right) \prod_{i=1}^{d} \psi'(u_i),$$

where $\psi_{(d)}$ is the d-th derivative of ψ. Two popular specifications of Archimedean copulas are the Gumbel copula and the Clayton copula with generators

$$\begin{aligned} \psi_\theta^{\text{G}}(x) &= \exp\left(-\log(x)^\theta\right), && x \in [0, \infty), \quad \theta \in [1, \infty), \\ \psi_\theta^{\text{C}}(x) &= (x^{-\theta} - 1)/\theta, && x \in [0, \infty), \quad \theta \in (-1/(d-1), \infty) \setminus \{0\}, \end{aligned}$$

and corresponding copulas

$$C_\theta^{\mathrm{G}}(x_1,\ldots,x_d) = \exp\left(-\left[(-\log(x_1))^\theta + \ldots + (-\log(x_d))^\theta\right]^{1/\theta}\right),$$
$$C_\theta^{\mathrm{C}}(x_1,\ldots,x_d) = \max\left\{x_1^\theta + \ldots + x_d^\theta - 1, 0\right\}^{-1/\theta}.$$

Gumbel copulas capture upper tail dependence while Clayton copulas capture lower tail dependence as illustrated in Figure 3.4.

3.4 Univariate GARCH Models

GARCH models are a generalization by Bollerslev (1986) of the autoregressive conditional heteroscedasticity (ARCH) model of Engle (1982). They are applied to a wide range of time series analysis and have been particularly successful in modelling financial returns (Engle, 2001). This is partly because the model captures volatility clustering effects which are often present in finance (Mandelbrot, 1963). In the presence of such effects, volatility becomes time-dependent and features aggregated periods of exceptionally high or low values.

The original vanilla GARCH(p, q) process describes the conditional variance as

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2, \tag{3.6}$$

where ε_t is the market shock or innovation at time t. The parameters $(\alpha_i)_{i=1}^q$ and $(\beta_i)_{i=1}^p$ measure the reaction of the conditional variance to market shocks and persistence of conditional variance respectively. To guarantee a positive long term variance, the model restricts the parameter choices to

$$\omega > 0, \qquad \alpha, \beta \geq 0, \qquad \alpha + \beta < 1. \tag{3.7}$$

For the market shock process, the model assumes

$$\varepsilon_t | \mathcal{I}_{t-1} \sim \mathcal{N}(0, \sigma_t^2), \tag{3.8}$$

where $\mathcal{I}_t$ is the information set containing all past returns up to t. The conditional variance is translated into the return through the conditional mean equation which in its simplest state is given by

$$r_t = c + \varepsilon_t, \qquad c \in \mathbb{R}. \tag{3.9}$$

Various extensions exist for both the conditional variance and conditional mean equation.[6] Most notably, the exponential GARCH (E-GARCH) is an asymmetric extension by Nelson (1991) that removes need for the parameter constraints of

[6] We refer to Teräsvirta (2009) for a survey of popular GARCH specifications.

Equation 3.7 by specifying the conditional variance equation in terms of log rather than directly. Given the asymmetric response function

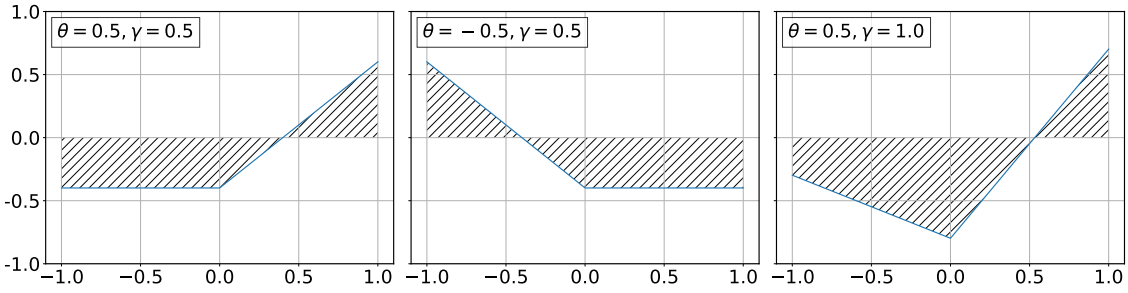
$$g(z_t) = \theta z_t + \gamma \left(|z_t| - \mathbb{E}(|z_t|) \right) \tag{3.10}$$

and $z_t = \varepsilon_t / \sigma_t$, the conditional variance and mean equations are

$$\begin{aligned} r &= c + \sigma_t z_t, \\ \log\left(\sigma_t^2\right) &= \omega + \sum_{i=1}^{q} \alpha_i g(z_{t-i}) + \sum_{i=1}^{p} \beta_i \log\left(\sigma_{t-i}^2\right). \end{aligned} \tag{3.11}$$

Positive and negative shocks can affect the variance differently, depending on the choices for γ and θ. This leverage effect is especially relevant in equity and commodity markets where the asymmetry is well documented. Figure 3.5 illustrates the great range of possible asymmetric responses in E-GARCH. The function g accounts for only positive shocks for $\theta = \gamma$, only negative shocks for $\theta = -\gamma$ and can model a large variety of reactions in between.

Figure 3.5: Asymmetric response function of E-GARCH for various parameters



The asymmetric response function g of Equation 3.10 is depicted with various parameters. We assume in this figure that z_t are Gaussian which means $\mathbb{E}(|z_t|) = \sqrt{2/\pi}$.

The variance modelled in Equations 3.6 and 3.11 is the variance of the distribution conditional on the current information set. As such it is time dependent and changes with additional historical data. In contrast, the unconditional variance of GARCH models is constant over time and usually interpreted as the long term average variance towards which the autoregressive process converges. If some prior estimate is available, GARCH can target the unconditional variance by fixing some of its parameters.

Forecasts for the returns are generated through an iterative process. Given a model calibrated on $t = 1, \ldots, T$ we use the last observed market shocks to estimate the future conditional variance $\hat{\sigma}_{T+1}$ which in turn gives access to $\hat{\varepsilon}_{T+1}$ through Equation 3.8.

The number of appropriate lags p, q can be determined by the Ljung-Box test (Ljung and Box, 1978). However, Hansen and Lunde (2005) compare various combinations of lag parameters and conclude that models with more lags rarely outperform the simple $p = q = 1$ benchmark even if the additional parameters are significant in-sample.

Maximum-likelihood estimation is used in conjunction with the assumption of Equation 3.8 to estimate the parameters. The normality assumption can be relaxed to allow for non-normal market returns with higher skewness or kurtosis. For instance, Bollerslev (1987) introduces Student-t distributed market shocks while more recent authors apply mixture normal distributions (Bai et al., 2003; Haas et al., 2004).

In practical applications, calibration of GARCH models may be difficult for several reasons:

(i) Optimization of the log-likelihood function can be challenging and should be done with advanced algorithms such as Levenberg-Marquardt. Alternatively, Markov Chain Monte Carlo (MCMC) may be applied. We refer to Virbickaite et al. (2015) for a review on the calibration of GARCH models with Bayesian methods.

(ii) The estimation of the unconditional covariance relies on a large calibration window, but even with large amounts of data the resulting estimate may not be accurate. Variance targeting with externally generated estimates for the long term variance can be applied to assist the estimation.

3.5 Multivariate GARCH Models

In multivariate analysis, clustering extends beyond volatilities to correlations. This motivates generalizations of univariate GARCH models to capture time varying conditional covariances in addition to the time varying conditional volatilities and to account for the spillover of volatility between the different assets.

Extending univariate GARCH models directly into higher dimensions is a challenge. The number of parameter estimations increases drastically with the dimension of the problem. Given d dimensions, $d(d+1)/2$ variances and covariances need to be estimated, each of which may depend on several parameters. Independent estimations also do not guarantee a positive definite correlation matrix. Furthermore, the likelihood curve is flat which may cause convergence errors through local maxima. We refer to Brooks et al. (2003) for a comparison of different multivariate GARCH implementations.

Bollerslev (1990) introduces the constant conditional correlation GARCH (CCC-GARCH) where the conditional correlations are assumed to be time-invariant. The covariance matrix is estimated as

$$\mathbf{V}_t = \mathbf{D}_t \mathbf{C} \mathbf{D}_t, \qquad \mathbf{D}_t = \operatorname{diag}\left(\mathbf{V}_t\right)^{1/2},$$

where $\mathbf{C}$ is a constant correlation matrix and $\mathbf{D}_t$ is the diagonal matrix containing the time-varying individual volatilities. The model is very easy to estimate since dependency and the volatilities are examined separately. Each volatility can be estimated by an univariate GARCH model while $\mathbf{C}$ can be specified as the sample correlation between standardized residuals. The number of parameters to estimate is in $\mathcal{O}(d)$. This in turn leads to a well-defined likelihood function which enables the use of CCC-GARCH in higher dimensions and guarantees a positive definite covariance matrix. However, the assumption of constant correlation may be too strong and is not fulfilled for many assets (Tsui and Yu, 1999).

Dynamic conditional correlation GARCH (DCC-GARCH) by Engle (2002) generalizes CCC-GARCH to account for time-varying but not stochastic correlations.

The correlation is estimated directly from the residuals of the univariate models and adjusted depending on the co-movement of the returns. As such, the covariance matrix is given by

$$\mathbf{V}_t = \mathbf{D}_t \mathbf{C}_t \mathbf{D}_t, \qquad \mathbf{D}_t = \operatorname{diag}\left(\mathbf{V}_t\right)^{1/2},$$

where the conditional correlation $\mathbf{C}_t$ with M and N lags is described by

$$\begin{aligned} \mathbf{C}_t &= \operatorname{diag}\left(\mathbf{Q}_t\right)^{-1/2} \mathbf{Q}_t \operatorname{diag}\left(\mathbf{Q}_t\right)^{-1/2}, \\ \mathbf{Q}_t &= \left(1 - \sum_{m=1}^{M} \alpha_m - \sum_{n=1}^{N} \beta_n\right) \overline{\mathbf{Q}} + \sum_{m=1}^{M} \alpha_m \left(\boldsymbol{\varepsilon}_{t-m} \boldsymbol{\varepsilon}_{t-m}'\right) + \sum_{n=1}^{N} \beta_n \mathbf{Q}_{t-n}, \end{aligned} \tag{3.12}$$

with

$$\overline{\mathbf{Q}} = \mathbb{E}\left(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t'\right).$$

The transformation of $\mathbf{Q}_t$ to $\mathbf{C}_t$ guarantees a well-defined correlation matrix as long $\mathbf{Q}_t$ is positive definite. Similar to CCC-GARCH, there are no restrictions on the choice of univariate GARCH models for the volatility. DCC-GARCH assumes normally distributed innovations but works if this assumption is not satisfied.[7] This is because the quasi-maximum-likelihood estimator remains consistent even if the distributional assumptions are incorrect (Bollerslev and Wooldridge, 1992; Cappiello et al., 2006). The number of parameters estimated during maximum likelihood remains $\mathcal{O}(d)$ as in CCC-GARCH but the dynamic correlation allows for easier interpretation and more flexible application. Cappiello et al. (2006) generalize the dynamic correlation in Equation 3.12 to allow for asymmetry but this increases the complexity to $\mathcal{O}(d^2)$ parameter estimations.

There are several alternative multivariate GARCH models. Bauwens et al. (2006) and Engle (2009, Chapter 3) survey the literature and list the following models as alternatives to CCC-GARCH and DCC-GARCH:

[7] There exist several extensions for DCC-GARCH to non-normal innovations. Bauwens and Laurent (2005) use multivariate skew distributions while Cajigas and Urga (2006) and Pelagatti (2004) apply Laplace distribution and elliptical distributions in general, respectively.

Diagonal vech GARCH The model by Bollerslev et al. (1988) describes each element of the covariance matrix by the product of the prior returns. Due to the generality, the model requires $\mathcal{O}(d^2)$ parameter estimations and might not yield positive-definite covariance matrices.

BEKK-GARCH Engle and Kroner (1995) adds restrictions to the diagonal vech GARCH by reducing the number of parameters and guarantees a positive definite covariance matrix. Despite the reduction in complexity, the model requires $\mathcal{O}(d^2)$ estimations.

Orthogonal GARCH Alexander (2002) uses a limited number of principal component factors to reduce the dimensionality of the covariance estimation. The model requires only $\mathrm{o}(d)$ estimations but might have poor performance in weakly correlated systems.

During our empirical analysis in Chapters 6 and 7, we limit our analysis to CCC-GARCH and DCC-GARCH for two reasons:

(i) Both models can be calibrated through a multi-step estimation procedure with comparatively low complexity. This allows us to analyse higher dimensional time series.

(ii) The models are successful in applications without relying on highly correlated returns. Engle and Sheppard (2008) compares various specifications of GARCH that are applicable in large systems and concludes that the DCC-GARCH family yields the best performance.

Chapter Four

Factor Quantile Methodology

4.1 Factor Quantile Regression 57
4.2 A Simple Example on Stock Returns 63
4.3 Factor Quantiles with Latent Factors 69
4.3.1 Alpha Quantile Forecasts 73
4.3.2 Bagging Quantile Forecasts 77

This chapter introduces Factor Quantile models, our new semi-parametric methodology for multivariate distribution forecasting where common factors describe each quantile of the dependent variables. Conditional marginal distributions are derived non-parametrically by quantile regression and combined into a conditional joint distribution through a parametric conditional copula. Factor Quantile models can be applied as a general forecasting method for a wide range of data sets and scale very well into higher dimensions due to the multi-stage approach.

Our literature review in Sections 2.2.1 and 2.2.2 lists several alternative forecasting methodologies with quantile regression. The favourable relative accuracy of these models against their respective benchmarks suggest that quantile regression can be successful in a forecasting setting. However, past models include many restrictions:

(i) Several studies use predictors that are unsuited for the forecasting problem. Cenesizoglu and Timmermann (2008) and Zhu (2013) apply forecast averaging to incorporate multivariate information into their single-factor models. However, in Cenesizoglu and Timmermann (2008), only 16% of the predictors are significant at 1%. Hence, it is unclear whether forecast averaging with equal weightings can yield appropriate estimates of the future quantiles, when forecasts are included that may be based on inadequate factor models. The empirical study of Zhu (2013) indicates similar issues, where only 9% of the factors for stock returns and 30% for bond returns are significant at 1%. In addition, some of the quantiles such as the median have no significant factor at all. Similarly, Gaglianone and Lima (2012) use forecasts of the expectation of the dependent variable to predict the future distribution function. This predictor may be unsuited, since it remains unclear why the expectation of a variable should contain information on other parts of its distribution.

(ii) Other models incorporate strong assumptions on the underlying data generating process or the availability of data. Ma and Pohlman (2008), for instance, assume the conditional location of their dependent variable to be constant over the forecasting period. Similarly, Gaglianone and Lima (2012) and Bunn et al.

(2016) rely on externally generated forecasts that may not be available in a general setting.

(iii) Some models rely on a large set of predictors which may be chosen ex-ante or through statistical variable selection methods (Manzan, 2015; Bunn et al., 2016; Meligkotsidou et al., 2019). However, the application of these models require a large amount of additional data and an understanding of the underlying process to specify the regression formula.

Section 4.1 starts with a discussion of the general idea of Factor Quantile models which uses a linear factor model to transform a point forecast of the common factors into a distribution forecast of the dependent variables. This model is contemporaneous and allows for the inclusion of multiple explanatory variables to describe the co-movement of the dependent variables. We illustrate all general concepts in Section 4.2 with a simple bivariate application on the daily stock returns of Apple and Procter & Gamble during the period 2000 – 2018. Following the basic methodology, we then present a latent factor version of our model in Section 4.3 where all factors are derived using endogenous principal component analysis and no external forecasts are required. Two different specifications of this latent model are described in Sections 4.3.1 and 4.3.2, based on the statistical properties of our principal component factors and bootstrap aggregation. Examples are provided with daily US interest rates from 1994 – 2018 for maturities between 6 months and 20 years.

4.1 Factor Quantile Regression

The starting point of our model description is a standard linear factor model

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \qquad t = 1, \dots, T,$$

where

$$\mathbf{y}_t = (\mathrm{y}_{1t}, \dots, \mathrm{y}_{dt})', \qquad \mathbf{x}_t = (\mathrm{x}_{1t}, \dots, \mathrm{x}_{mt})',$$

denote the time t values of d dependent variables and m common factors. We set $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)'$ as the constant vector of intercepts, $\mathbf{B}$ as the constant matrix of factor sensitivities, and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \dots, \varepsilon_{dt})'$ as the vector of error processes. Further, we assume that the observations $(\mathbf{y}_t)_{t=1}^T$ arise from a random variable $\mathbf{Y} = (Y_1, \dots, Y_d)'$ with stationary conditional joint distribution $F|\mathbf{x}_t$ and conditional marginal distributions

$$F_1|\mathbf{x}_t, \dots, F_d|\mathbf{x}_t.$$

Similar linear models with common macroeconomic, fundamental or statistical factors have been introduced by Ross (1976), Fama and French (1993) and Connor et al. (2012) respectively and are well established in several areas of finance and economics. Applications include portfolio management (Ross, 1976; Fama and French, 1993; Connor et al., 2012), risk analysis (Avramidis and Pasiouras, 2015; Bunn et al., 2016; Chou et al., 2017; Tu and Chen, 2018) and forecasting with economic policy implications (Patton, 2006; Duan and Miao, 2016; Coroneo et al., 2016; Kavtaradze and Mokhtari, 2018; Wellmann and Trück, 2018; Cheung et al., 2018). The main focus of such factor models is to attribute the variance in the dependent variables to different common factors that are treated as independent variables. However, standard estimation techniques such as ordinary least squares are limited to inferences on the means and variances of the dependent variables, conditional on each factor.

By contrast, factor quantile regressions allow the explanatory variables to affect the dependent variables differently for each τ-quantile, and estimation can trace out

the conditional distribution of each dependent variable as τ ranges from 0 to 1. We extend the contemporaneous quantile-regression framework of Gaglianone and Lima (2012) to multiple factors and capture this flexibility as

$$\mathbf{y}_t^{(\tau)} = \boldsymbol{\alpha}^{(\tau)} + \mathbf{B}^{(\tau)}\mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(\tau)}, \qquad t = 1, \ldots, T, \tag{4.1}$$

with a quantile-dependent error process $\boldsymbol{\varepsilon}_t^{(\tau)}$, as well as constants for the intercept $\boldsymbol{\alpha}^{(\tau)}$ and matrix of quantile regression coefficients $\mathbf{B}^{(\tau)}$. The regressand $\hat{\mathbf{y}}_t^{(\tau)}$ estimates

$$(Q_{Y_1}(\tau|\mathbf{x}_t), \ldots, Q_{Y_d}(\tau|\mathbf{x}_t))'$$

and represents the vector containing the τ-quantile of each element of $\mathbf{y}_t$, conditional on $\mathbf{x}_t$.

The contemporaneous relationship between dependent and explanatory variables in our Factor Quantile model is motivated by the generally weak fit of forecasting models with lagged explanatory variables especially when multiple quantiles are considered. In the studies of Cenesizoglu and Timmermann (2008) and Zhu (2013), most of the lagged economic predictors for the stock and bond returns are not statistically significant in the quantile regressions. In contrast, Bunn et al. (2016) utilize contemporaneous information in their quantile model which performs well against asymmetric GARCH models with non-normal innovations.

Therefore, to derive conditional forecasts for our dependent variables, the explanatory variables need to be predicted. In general, Factor Quantile models may use any externally-generated forecast $\hat{\mathbf{x}}_{T+1}$ which considers the dependency structure between the explanatory variables. Assuming such a forecast is available we can estimate the quantile regressions using historical data for $t = 1, \ldots, T$, and then predict each conditional quantile at time $T + 1$ as

$$\hat{\mathbf{y}}_{T+1}^{(\tau)} = \boldsymbol{\alpha}^{(\tau)} + \mathbf{B}^{(\tau)}\hat{\mathbf{x}}_{T+1}.$$

Next consider a quantile partition $\mathbb{Q}$ where $0 < \tau < 1$ for all $\tau \in \mathbb{Q}$ and focus for now on the i-th element of $\mathbf{y}_t$. If $\mathbb{Q}$ outlines a sufficiently dense grid, the shape

of the entire forecasted conditional distribution function $F_i|\mathbf{x}_{T+1}$ of $\mathrm{y}_{i,T+1}$ can be estimated through

$$\left\{\left(\tau, \hat{\mathrm{y}}_{i,T+1}^{(\tau)}\right) : \tau \in \mathbb{Q}\right\}.$$

The optimal node positions depend on $F_i|\mathbf{x}_{T+1}$ and should focus more on parts where the distribution is expected to be irregular. Since fitting the tails of the distribution is more of a challenge than fitting the centre, nodes concentrated around the tails are beneficial.

Multiple methods have been applied to interpolate a continuous distribution from the estimated quantiles:

(i) Koenker and Bassett (1982) use a step function which assigns the value of the next smallest quantile in $\tau \in \mathbb{Q}$. This method is adapted by Cenesizoglu and Timmermann (2008) and Pedersen (2015);

(ii) Kernel density estimations, e.g. with Gaussian or Epanechnikov kernel, can be employed as in Koenker and Bassett (2010) and Gaglianone and Lima (2012).

Instead of those established methods, we estimate the distribution through interpolation with cubic Hermite splines since this yields a well-defined distribution estimate that is efficient to compute – see Section 4.2 for a more detailed comparison.

Definition 4.1 (Cubic Hermite spline)**.** Let $f : [a, b] \to \mathbb{R}$ be an unknown function going through $(x_1, f(x_1)), \ldots, (x_n, f(x_n))$ with

$$a = x_1 < x_2 < \ldots < x_n = b$$

and with slopes $f'(x_1), \ldots, f'(x_n)$. Define the Hermite basis functions

$$\begin{aligned} h_{00}(x) &= (1 + 2x)(1 - x)^2, & h_{01}(x) &= x^2(3 - 2x), \\ h_{10}(x) &= x(1 - x)^2, & h_{11}(x) &= x^2(x - 1). \end{aligned}$$

and let $z_i(x) = (x - x_i)/(x_{i+1} - x_i)$. Then, the cubic Hermite spline is given by

$$\hat{f}(x) = \begin{cases} f(x) & \text{for } x \in \{x_1, \ldots, x_n\} \\ h_{00}\left(z_i(x)\right) f(x_i) & \text{for } x_i < x < x_{i+1} \\ \quad +h_{01}\left(z_i(x)\right) f(x_{i+1}) & \\ \quad +h_{10}\left(z_i(x)\right)\left(x_{i+1} - x_i\right) f'(x_i) & \\ \quad +h_{11}\left(z_i(x)\right)\left(x_{i+1} - x_i\right) f'(x_{i+1}) & \end{cases} .$$

Definition 4.1 yields a unique third-degree polynomial path with the given points and slopes. There are several algorithms to estimate the slopes at the interpolation points $f'(x_1), \ldots, f'(x_n)$:

(i) Akima (1970) uses a method similar to geometric means;

(ii) Ellis and McLain (1977) apply a least squares procedure;

(iii) Passow (1974) sets $f'(x) = 0$ for all $x \in \{x_1, \ldots, x_n\}$;

(iv) Alternatively, the slope can approximated by a two-point formula

$$f'(x_i) = \begin{cases} \left(f(x_{i+1}) - f(x_i)\right)/(x_{i+1} - x_i) & \text{for } i \in \{1, \ldots, n-1\} \\ 0 & \text{for } i = n. \end{cases}$$

We prefer the slopes of Fritsch and Carlson (1980) which result in a piecewise cubic Hermite interpolating polynomial (PCHIP) that is continuously differentiable and preserves the monotonicity in the estimated quantiles. Thereby, our distribution and density function estimates are well-defined. Unlike kernels, it imposes no assumptions about the shape and maintains the original shape well even if $\mathbb{Q}$ has low cardinality. Section 4.2 elaborates the advantages of the shape preserving interpolation by comparing the effectiveness and efficiency of various estimation methods for stock return data.

Given a forecast $\hat{\mathbf{x}}_{T+1}$ of the common factors, denote the interpolated conditional distribution functions by

$$\hat{F}_i|\hat{\mathbf{x}}_{T+1}, \qquad \text{for } i = 1, \ldots, d.$$

The probability integral transform variables are uniformly distributed if the forecast is probabilistically calibrated and will only be independent if the residuals

$$\varepsilon_{i,T+1}|\hat{\mathbf{x}}_{T+1} = F_i - \hat{F}_i|\hat{\mathbf{x}}_{T+1}$$

are independent which may be not the case unless the factor model perfectly represents the regressand without any missing variables or similar problems. Otherwise, we capture dependence using an extension of Sklar's theorem to conditional copulas due to Patton (2006) which represents a joint conditional distribution in terms of a unique conditional copula defined by

$$\hat{F}\left(\mathbf{y}|\hat{\mathbf{x}}_{T+1}\right) = C\left(\hat{F}_1\left(y_1|\hat{\mathbf{x}}_{T+1}\right), \ldots, \hat{F}_d\left(y_d|\hat{\mathbf{x}}_{T+1}\right)\middle|\, \hat{\mathbf{x}}_{T+1}\right). \tag{4.2}$$

This way, any conditional marginals can be transformed into a valid multivariate distribution provided the copula is conditioned on the same variables as the marginal distributions. As Patton (2013) points out, this multi-stage approach results in a multivariate model without the challenges associated with simultaneous estimations in high dimensions.

To summarize, the general methodology of Factor Quantile models proceeds as follows:

Stage 1 Estimate quantile regressions for τ-quantiles where $\tau \in (0, 1)$ are pre-specified by a partition $\mathbb{Q}$;

Stage 2 For a given vector $\hat{\mathbf{x}}_{T+1}$ for the common factors, interpolate over conditional quantiles in $\mathbb{Q}$ to obtain each conditional marginal $\hat{F}_1|\hat{\mathbf{x}}_{T+1} \ldots, \hat{F}_d|\hat{\mathbf{x}}_{T+1}$;

Stage 3 Use a conditional copula and apply Equation 4.2 to obtain the joint conditional distribution.

Algorithm 1 summarizes the pseudo-code in a d-dimensional distribution forecasting setting.

Algorithm 1: Factor Quantile model

Input : Factor model from Equation 4.1 and conditional copula C;
Quantile partition $\mathbb{Q}$ with $0 < \tau < 1$ for all $\tau \in \mathbb{Q}$;
Observations on $\mathbf{y}_t$ and $\mathbf{x}_t$ for $t = 1, \ldots, T$;
Externally generated forecast $\hat{\mathbf{x}}_{T+1}$;

Output : Conditional multivariate distribution $\hat{F}|\hat{\mathbf{x}}_{T+1}$ of $\mathbf{y}_t$;

for $i = 1, \ldots, d$ **do**
  Use historical data $t = 1, \ldots, T$ to estimate the factor quantile regressions using $\boldsymbol{\beta}_i^{(\tau)}$, the i-th row of $\mathbf{B}^{(\tau)}$:

$$\mathrm{y}_{it}^{(\tau)} \leftarrow \alpha_i^{(\tau)} + \boldsymbol{\beta}_i^{(\tau)} \mathbf{x}_t + \varepsilon_{it}^{(\tau)}$$

  which yields $\hat{\alpha}_i^{(\tau)}$ and $\hat{\boldsymbol{\beta}}_i^{(\tau)}$ for each $\tau \in \mathbb{Q}$;
  Use the externally generated forecast $\hat{\mathbf{x}}_{T+1}$ to compute conditional quantile forecasts

$$\hat{\mathrm{y}}_{i,T+1}^{(\tau)} \leftarrow \hat{\alpha}_i^{(\tau)} + \hat{\boldsymbol{\beta}}_i^{(\tau)} \hat{\mathbf{x}}_{T+1}, \qquad \tau \in \mathbb{Q};$$

  Estimate $\hat{F}_i|\hat{\mathbf{x}}_{T+1}$, the conditional distribution function of $\mathrm{y}_{i,T+1}$, through shape-preserving interpolation on

$$\left\{ \left(\tau, \hat{\mathrm{y}}_{i,T+1}^{(\tau)} \right) : \tau \in \mathbb{Q} \right\};$$

end
Generate the conditional multivariate distribution with the marginal distributions and a conditional copula

$$\hat{F}(\mathbf{y}|\hat{\mathbf{x}}_{T+1}) \leftarrow C\left(\hat{F}_1(y_1|\hat{\mathbf{x}}_{T+1}), \ldots, \hat{F}_d(y_d|\hat{\mathbf{x}}_{T+1}) \middle| \hat{\mathbf{x}}_{T+1} \right);$$

4.2 A Simple Example on Stock Returns

We illustrate the general Factor Quantile model in the case where dependent variables are excess stock returns $r_{1t}, \ldots, r_{dt}$ and the factor model is the two-factor Capital Asset Pricing Model (CAPM) introduced by Kraus and Litzenberger (1976). Through the inclusion of a quadratic term in the excess market return r_M, the two-factor CAPM captures different sensitivities to positive and negative returns and allows the systematic risk of a stock to be related to skewness, as in Harvey and Siddique (2000). Throughout this chapter, we assume that the risk-free interest rate is zero so that return and excess return are equal. This is justified by our focus on the general Factor Quantile model methodology rather any specific regression model.

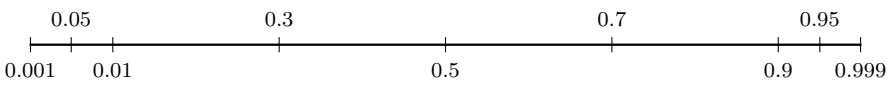
The quantile regressions for the i-th stock return may be written as

$$r_{it}^{(\tau)} = \alpha^{(\tau)} + \beta^{(\tau)} r_{tM} + \gamma^{(\tau)} r_{tM}^2 + \varepsilon_{it}^{(\tau)}, \qquad t = 1, \ldots, T. \tag{4.3}$$

For simplicity of the graphical representations, we limit our discussion in this example to the bivariate case $d = 2$.

First we consider the selection of the quantile partition and compare the properties of the three interpolation methods described in Section 4.1. To this end, we estimate quantile regressions for returns on the stock Apple with the S&P500 as market factor and two different quantile partitions $\mathbb{Q}_9$ and $\mathbb{Q}_{500}$, where $|\mathbb{Q}_9| = 9$ and $|\mathbb{Q}_{500}| = 500$. The larger quantile partition utilizes equidistant nodes which cover $(0, 1)$ in a dense grid. With $|\mathbb{Q}| = 9$ we add more nodes in the extremes to better capture the tail behaviour:

$$\mathbb{Q}_9 = \{0.001, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.999\}$$

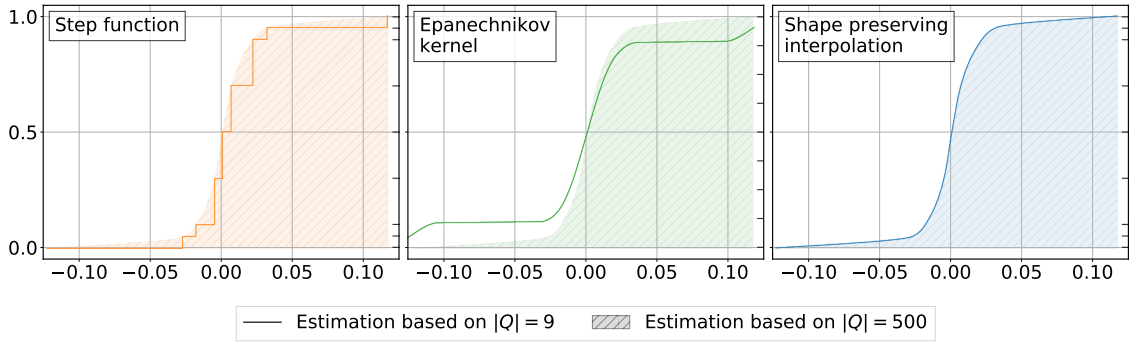


(4.4)

However, quantile regression is likely to yield high sampling error for the extreme nodes because there are fewer data points in those percentiles, by definition. But, on balance, taking account of the monotonicity requirement for quantiles and the

hit-or-miss accuracy of ad-hoc extrapolation, additional nodes in the tails should benefit the accuracy of the estimated distribution nevertheless. Figure 4.1 compares the results for (i) the step function introduced by Koenker and Bassett Jr (1978) on the left in orange; (ii) the Epanechnikov kernel advocated by Gaglianone and Lima (2012) in the middle in green;[1] and (iii) the shape-preserving interpolation on the right in blue.

Figure 4.1: Distribution estimates with varying quantile partitions (Apple)



Conditional distributions for the return on Apple based on an equidistant quantile partition $\mathbb{Q}_{500}$ with $|\mathbb{Q}| = 500$ (shaded area) are compared with distributions based on $|\mathbb{Q}| = 9$ (solid line). The step function and the shape-preserving interpolation utilize $\mathbb{Q}_9$ with a focus on the tails while the kernel estimation uses equidistant nodes as illustrated with the rugs on the right-side axis. All conditional quantiles are based on the quadratic CAPM in Equation 4.3 and are calibrated on data from 03 January 2000 to 28 June 2018. The market return is on the S&P500 index and all distributions are conditional on the realized S&P return on 29 June 2018.

The quantile partition $\mathbb{Q}_{500}$ produces very similar distributions for all three methods which are indistinguishable in a Kolmogorov-Smirnov test at significance level of 1%. However, with $|\mathbb{Q}| = 9$ the shape-preserving interpolation fits much better than the kernel or the step function, the latter two yielding vastly different distributions depending on the choice of $\mathbb{Q}$.

To quantify the additional quantile partition requirements of the kernel and the step function, we sample from distributions with equidistant quantile partitions of varying cardinality and compare them with the estimation based on $\mathbb{Q}_{500}$ through a Kolmogorov-Smirnov test in Table 4.1. The kernel requires $|\mathbb{Q}| = 35$ and the step function $|\mathbb{Q}| = 50$ to achieve a similar distribution. However, the shape-preserving interpolation with $|\mathbb{Q}| = 9$ yields a function which a Kolmogorov-Smirnov test cannot

[1] To facilitate a fair comparison, the kernel uses an equidistant quantile partition even in the case with only few nodes since this yields a better distribution estimate.

distinguish from the one based on $\mathbb{Q}_{500}$ at a significance level of 1%. The lower cardinality requirement of the shape-preserving interpolation is especially relevant in practice since it leads to major computational improvements. The total time taken for estimating all quantile regressions and then applying the distribution estimation with $|\mathbb{Q}| = 9$, 35 and 50, respectively, is over four times longer for both the kernel and the step function than the shape preserving interpolation.[2]

Table 4.1: Kolmogorov-Smirnov p-values of distribution comparison (Apple)

$\lvert\mathbb{Q}\rvert$	Step function	Epanechnikov kernel
10	0.0027	0.2562
20	0.4493	0.9154
30	0.8110	0.9855
40	0.9885	0.9996
50	0.9997	0.9996

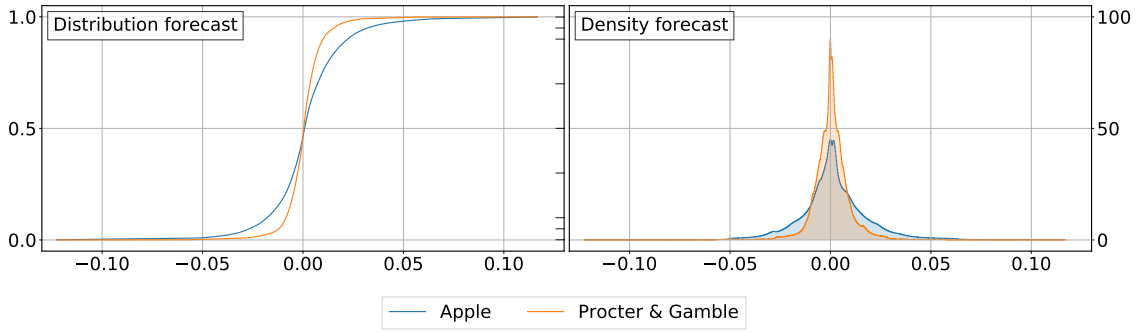
The quantiles for the return of Apple are calculated with the quadratic CAPM in Equation 4.3 and data from 03 January 2000 to 28 June 2018. We model the market return through the returns of the S&P500 index and condition all distributions on the realized S&P return from 29 June 2018.

Next we estimate quantile regressions Equation 4.3 on another US stock, Procter and Gamble (P&G) over the same time period. Interpolating allows for a visual comparison of the conditional distributions and densities of Apple and P&G, depicted in Figure 4.2. During the period 2000 – 2018 Apple returns were highly volatile, as is evident from the broader range of support for the Apple density and the steeper slope of the distribution for P&G. Both distributions and densities are smooth and exhibit irregularities which are difficult to capture with alternative parametric estimations.

Now we use these conditional marginal distributions to illustrate our Factor Quantile model based on a bivariate copula by fitting the conditional joint distribution with a Gaussian, Gumbel and Clayton copula. Table 4.2 summarizes the goodness of fit which identifies the Gumbel copula as the most suitable choice for our data. The conditional joint density forecasts are illustrated in Figure 4.3 which show slight but noticeable differences depending on the copula choice.

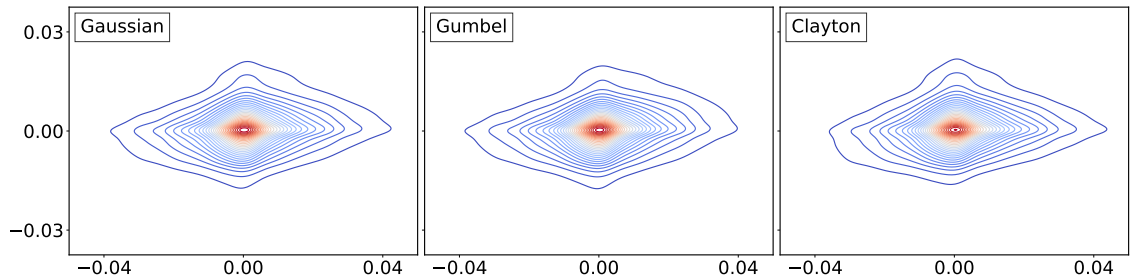
[2]Using an Intel i5-6500 with 3.20 GHz, shape preserving interpolation needs 475 ± 11 ms, while kernel and step function require 1,910 ± 63 ms and 4,080 ± 207 ms.

Figure 4.2: Conditional distribution and density forecasts (Apple and P&G)



The conditional marginal distribution and corresponding density for two US stock returns are generated with a Factor Quantile model based on the quadratic CAPM in Equation 4.3. For the calibration, we use data from 03 January 2000 to 28 June 2018 as well as a quantile partition $\mathbb{Q}_9$ as illustrated with the rugs on the right-side axis. The market return is on the S&P500 index and both distributions are conditional on the realized S&P return from 29 June 2018.

Figure 4.3: Joint conditional density forecasts (Apple and P&G)



We use maximum likelihood estimation on the stock returns from 03 January 2000 to 28 June 2018 to derive the optimal parameters for the Gaussian and Archimedean copulas. This yields $\rho = 0.1988$ for the Gaussian copula and $\theta = 1.1590$ or $\theta = 0.2690$ for the Gumbel and Clayton copula respectively.

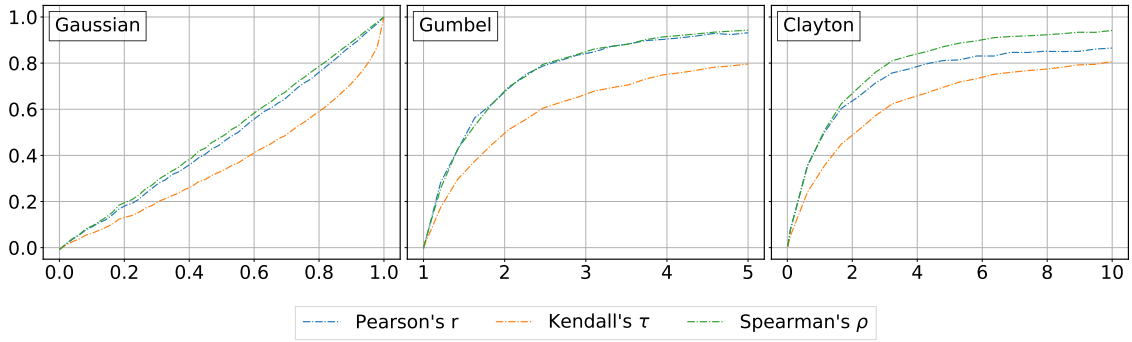
Table 4.2: Information criteria for the copula fit (Apple and P&G)

Criterion	Gaussian	Gumbel	Clayton
$\log(\hat{L})$	96.68	130.75	109.14
AIC	-216.28	-259.50	-191.36
BIC	-214.60	-257.82	-189.68

All copula models use only one parameter since we consider a bivariate case. For the calibration, we use stock returns from 03 January 2000 to 28 June 2018. The definition of the information criteria can be found in Section 3.3. A high log likelihood or a low information criteria suggests a good fit.

Our semi-parametric Factor Quantile model allows for a wide variety of correlation and dependency structures. Figure 4.4 compares the values of the standard Pearson correlation and the two standard rank correlation metrics Kendall's τ and Spearman's ρ as the parameter of each copula varies.[3] These figures illustrate how a target correlation – such as may be applied in stress testing the two-stock portfolio – can be transformed into a unique value for the copula parameter which can be used in the Factor Quantile algorithm.

Figure 4.4: Joint conditional density forecasts (Apple and P&G)



The correlation measures are calculated by simulating from a bivariate distribution based on the conditional joint distributions of Figure 4.3. On the x-axis is the parameter for the respective copula, namely Pearson's correlation for the Gaussian copula and Archimedean θ for the Gumbel and Clayton copula. The samples are created using rejection sampling. Slight irregularities and non-monotonicity arises from simulation error and could be reduced by increasing the simulation size.

We should emphasize that the entire dependency structure between the conditional marginal distributions of Factor Quantile models are derived from the choice and

[3]The relationship between the Archimedean copula parameter θ and Kendall's τ is known analytically for both the Gumbel and the Clayton copula as $\tau_G = 1 - \theta^{-1}$ and $\tau_C = \theta/(\theta + 2)$ respectively. However, the association with Spearman's rho is not available in a simple form and there is no formula governing the relation with Pearson's correlation since the latter depends on both the marginals and the copula.

parametrisation of the conditional copula. The regression models for the conditional quantile forecasts in Equation 4.1 share the same predictor variables $\mathbf{x}_t$, but this does not affect conditional rank correlation metrics such as Kendall's τ or Spearman's ρ. Of course, the unconditional dependency depends on both the copula and the factor structure since the movement of $\mathbf{x}_t$ affects all dependent variables simultaneously. Hence, one way to pick a target correlation for the copula portion from historical data is to (i) calibrate the regression model for each element of $\mathbf{y}_t$ with ordinary least squares (OLS) and then (ii) calculates the conditional correlation through the OLS residuals.

4.3 Factor Quantiles with Latent Factors

Now consider the case that common factors are latent variables corresponding to principal components of the covariance matrix of $\mathbf{y}_t$.[4] This generalizes our methodology to allow for its application when no suitable factor models or externally generated forecasts of the common factors are available.

Following Stock and Watson (2002), many papers on quantile regression employ principal components derived from the covariance matrix of a set of exogenous predictor variables. For example, Ando and Tsay (2011) explore theoretical properties of quantile regression models with explanatory variables that include such principal components, developing an information-theoretic criterion to determine the optimal number of components to include. Manzan (2015) empirically evaluates the predictive power of principal components of a large number of exogenous macroeconomic indicators when used to augment the Koenker and Xiao (2006) autoregressive model for quantiles. Maciejowska et al. (2016) generalize the quantile regression averaging approach by Nowotarski and Weron (2015) with principal components to avoid the ex-ante model selection. Quantile regression averaging involves applying quantile regression with a set of individual point forecasts as independent variables and the observed value of the predicted variable as the dependent variable.

By contrast, we are interested in the case that the latent factors are endogenous, in the sense that the principal components are derived from the covariance matrix of the dependent variables alone.[5] This endogenous approach was first employed by Connor and Korajczyk (1993) who use asymptotic results on principal components to determine the appropriate number of factors for explaining returns on US stocks.

Given observations $(\mathbf{y}_t)_{t=1}^T$ of the dependent variables, denote the matrix of eigenvectors of the sample covariance matrix $\mathbf{V}$ by $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_d)$. Order the

[4]Principal components here are defined as time series contrary to their introduction as random variables in Section 3.2. However, since the time series observations can be interpreted as realizations from some random variable, all prior results remain valid.

[5]To differentiate this from macroeconomic or fundamental (e.g. Fama and French (1993) type models) of financial asset returns, Connor (1995) calls this a statistical factor model.

columns of $\mathbf{W}$ so that $\mathbf{w}_i$ is the eigenvector corresponding to λ_i, the i-th largest eigenvalue of $\mathbf{V}$. Set

$$\mathbf{p}_t = (p_{1t}, \ldots, p_{dt})' = \mathbf{W}'\mathbf{y}_t$$

so that p_{it} is the i-th principal component at time t. Because it is orthogonal, $\mathbf{W}' = \mathbf{W}^{-1}$, so the principal component representation is

$$\mathbf{y}_t = \mathbf{W}\mathbf{p}_t$$

as discussed in Section 3.2. Then a statistical factor model, based on endogenous principal component factors, is an approximate representation

$$\mathbf{y}_t \approx \mathbf{W}^m \mathbf{x}_t^m,$$

where $\mathbf{W}^m = (\mathbf{w}_1, \ldots, \mathbf{w}_m)$ denotes the first m columns of $\mathbf{W}$ and

$$\mathbf{x}_t^m = (\mathrm{p}_{1t}, \ldots, \mathrm{p}_{mt})'.$$

The approximation is justified by the decreasing amount of variance explained of the higher principal components and maximizes the variance explained amongst any linear representation with m factors.

We select the number of factors m so that a large fraction of the total variance is explained and the amount of unwanted noise which is not useful for forecasting is limited. Typically we choose m to explain around 90% or 95% of variation, regarding the remaining 5% or 10% as noise. This way the errors in an ordinary multiple regression of $\mathbf{y}_t$ on $\mathbf{x}^m$ would have very small variances and covariances. Indeed, in-sample point estimates for each dependent variable may be derived as

$$\hat{\mathbf{y}}_t = \mathbf{W}^m \mathbf{x}_t^m$$

without needing ordinary least squares. Similarly, given point forecasts $\hat{\mathbf{x}}_{T+1}^m$ for the principal components we may set

$$\hat{\mathbf{y}}_{T+1} = \mathbf{W}^m \hat{\mathbf{x}}_{T+1}^m,$$

which adjusts the quantile regression in Equation 4.1 to

$$\hat{\mathbf{y}}_t^{(\tau)} = \boldsymbol{\alpha}^{(\tau)} + \mathbf{B}^{(\tau)}\mathbf{x}_t^m + \boldsymbol{\varepsilon}_t^{(\tau)}, \qquad t = 1, \ldots, T. \tag{4.5}$$

This statistical factor model has a high R^2 as long as the number of principal components are chosen such that the total variance explained is high. As discussed in Section 2.2.3, the distribution estimate from those conditional quantile forecasts outperform alternative, direct estimations of the distribution function (Koenker and Leorato, 2015).

Applying our latent factor model to zero-expectation regressors, as is usually the case with returns in financial and economic data sets, or by centering the principal components, we can further set

$$\mathbb{E}\left(\mathbf{x}_t^m\right) = \mathbf{0}.$$

Although generally $\mathbb{E}(\boldsymbol{\varepsilon}_t^{(\tau)}) \neq 0$, when m is sufficiently large the errors in Equation 4.5 are small enough to be ignored. Therefore, we can write the expectation of each conditional quantile as

$$\mathbb{E}\left(\hat{\mathbf{y}}_t^{(\tau)}\right) = \boldsymbol{\alpha}^{(\tau)}. \tag{4.6}$$

Further, since the principal components are uncorrelated to each other, the variance of the τ-quantile is given by

$$\mathbb{V}\mathrm{ar}\left(\hat{\mathbf{y}}_t^{(\tau)}\right) = \mathbf{B}^{(\tau)}\mathrm{diag}(\lambda_1, \ldots, \lambda_m)\mathbf{B}^{(\tau)\prime}. \tag{4.7}$$

Similarly, we get the analytical form for the the covariance between some τ_1- and τ_2-quantile as

$$\mathbb{C}\mathrm{ov}\left(\hat{\mathbf{y}}_t^{(\tau_1)}, \hat{\mathbf{y}}_t^{(\tau_2)}\right) = \mathbf{B}^{(\tau_1)}\mathrm{diag}(\lambda_1, \ldots, \lambda_m)\mathbf{B}^{(\tau_2)\prime}. \tag{4.8}$$

In Section 4.3.1 and 4.3.2, we will use the fact that the expectation, variance and covariance matrix are not time-dependent because we assume the observations $\mathbf{y}_t$ to arise from a stationary conditional joint distribution.

Our latent factor model has several advantages over other macroeconomic or fundamental ones:

(i) The principal component representation is a valid linear model that works irrespective of choice of dependent variables $\mathbf{y}_t$ and therefore remains applicable even when other factor models are difficult to obtain. Especially in applications of higher dimensions, the dependent variables often have a correlation structure that facilitates the use of PCA. The principal component representation does not work well for weakly correlated systems, but this in turn may mean that the joint distribution becomes less interesting in general.

(ii) We have flexibility to select the amount of variance explained by the factor model and by doing so we limit the noise captured.

(iii) The use of principal components leads to robust estimates since all factors are uncorrelated with each other.

Contrary to cases where we are interested in the determinants of multivariate systems, we are not affected by the lack of interpretability of the principal components.

In the subsequent sections, we introduce several specifications of Factor Quantile models utilizing the principal component representation in Equation 4.5. Examples on US interest rate changes illustrate each approach.

4.3.1 Alpha Quantile Forecasts

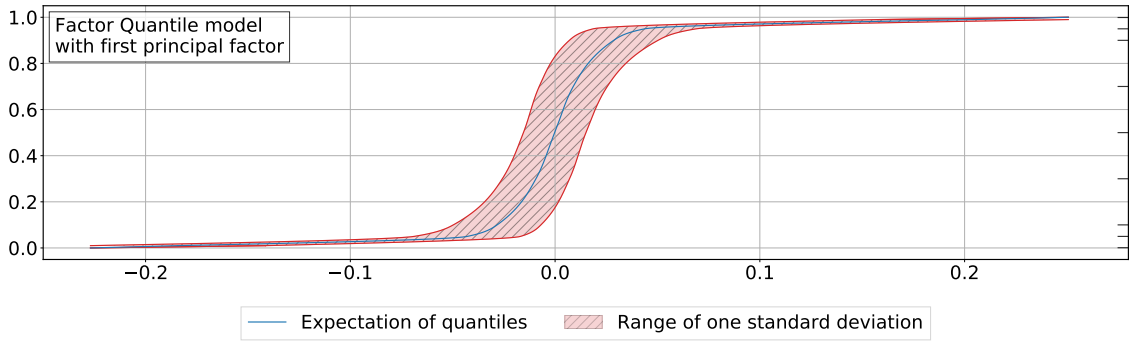
Based on Equation 4.5, one straightforward linear factor model is the principal component representation with the first m principal components. All conditional quantiles in a quantile partition $\mathbb{Q}$ can be forecasted by their expectation

$$\hat{\mathbf{y}}_{T+1}^{(\tau)} = \mathbb{E}\left(\mathbf{y}_{T+1}^{(\tau)}\right) = \boldsymbol{\alpha}^{(\tau)}$$

as outlined in Equation 4.6. There are no issues with quantile crossing since the quantile loadings are calculated conditional on the expectation of the explanatory variables. We denote these type of distributions which use solely the expectation of the future quantiles as Factor Quantile Alpha distributions.

Unfortunately, this naïve model is unsuited to forecasting. The variance around each quantile forecast described in Equation 4.7 is considerable since the first principal components contain the most variance explained and belong to relatively large eigenvalues. Therefore, even in a model with few components, each quantile forecast is associated with a large uncertainty.

Figure 4.5: Forecast with first principal factors (6 month interest rate)



The conditional marginal distribution for the 6 month US interest rate changes is generated with a Factor Quantile model based on the principal component representation in Equation 4.5 and the first principal component. We use data from 03 January 1994 to 29 June 2018 and the quantile partition of Equation 4.4 for the calibration as illustrated with the rugs on the right-side axis.

Figure 4.5 illustrates this issue with US interest rate data and a linear factor model with one principal factor.[6] This representation explains 84% of the original variation as the data is highly correlated. The distribution forecast resulting from

[6]The principal component representation is based on daily changes in interest rates of 6 month, 1 year, 2 year, 3 year, 5 year, 7 year, 10 year and 20 year maturity.

Equation 4.5 is depicted in blue with a surrounding red area which covers the range of one standard deviation around the expectation of each quantile forecast. Given the wide interval around the expectation, it is unclear whether the distribution is an adequate forecast since there are many alternatives that are also likely but may deviate strongly from the originally proposed prediction.

There are two reasons for the poor forecasting suitability of the principal representation model with the first few principal components. First, by utilizing the zero expectation of the principal factors, we disregard the quantile loadings attributed to the explanatory variables. Only variation captured in the intercept $\boldsymbol{\alpha}^{(\tau)}$ is deployed in the forecast which might not be enough to yield a good estimate. Second, each quantile has a large variance in Equation 4.7 which makes the expectation by itself insufficient for accurate predictions. Any estimate near the expectation may also be associated with a high probability of realization since the exact distribution of each quantile is unknown.

A better alternative to this naïve model is one that uses the last few principal components

$$\mathbf{x}_t^m = (\mathrm{p}_{dt}, \ldots, \mathrm{p}_{d-m+1,t})'$$

instead in Equation 4.5. By regressing on the last principal components, we separate the relevant information and the noise, which are captured by the intercept $\boldsymbol{\alpha}^{(\tau)}$ and the quantile loadings respectively:

(i) We interpret the variation captured by the last principal components as noise. Conversely, we want to retain all variation that is not captured by the last principal components;

(ii) During our regression, we encompass all variance that cannot be explained by the last few principal components in the intercept (and in the error).

Therefore, our factor model with the last principal components removes unwanted noise and reduces the variance of the quantiles through the (constant) intercept.

Statistical properties described in Equations 4.6, 4.7 and 4.8 remain valid because our common factors are still uncorrelated principal components with zero expectation. Hence, future quantiles can be approximated by their expectation $\boldsymbol{\alpha}^{(\tau)}$, which incorporates any variation that cannot be explained by the last principal components. As before, there are no difficulties arising from quantile crossing because we condition on the expectation of the explanatory variables.

Simultaneously, the variation around the expectation of the quantiles is reduced greatly. The last principal components have the smallest eigenvalues of all principal components and further, since they explain the least amount of the variance, it is likely that their factor loadings are smaller than those for the first few principal components in the naïve model. This leads to a lower variance for the conditional quantiles through Equation 4.7.

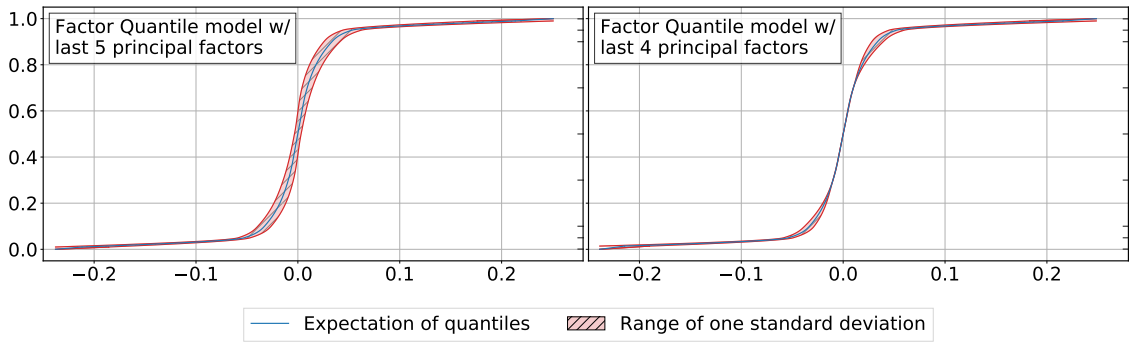
The concept of using the intercept to encompass the remaining variation not explained by the factors in the linear regression is widely applied in performance evaluation of portfolio managers, following the introduction of Jensen's Alpha by Jensen (1968). This measure is among the most widely used performance metrics and can be utilized with many regression models under general assumptions (Goetzmann et al., 2007).[7]

Figure 4.6 shows the distributions based on regressions with the last five and four principal components which explain 3% and 2% of the variance. The range of one standard deviation is much smaller compared to the distribution based on the first principal component. Although both forecasts are similar to the one from Figure 4.5 in this example because they refer to the same distribution, this does not need to be the case generally. A Kolmogorov-Smirnov test distinguishes between the distribution in Figure 4.5 and the ones in Figure 4.6 at a significance level of 1%.

Given the magnitude of the variance of the future quantiles, the approach with the last few principal components seems to be the better choice than the one with the

[7]Most of the performance evaluation literature criticising Jensen's Alpha focus on misspecified regression models, where a positive intercept may be due to omitted variables rather than abnormal performance of a fund manager – see Jarrow and Protter (2013). However, since we are mostly interested in separating the relevant information and the noise, this does not affect us.

Figure 4.6: Forecast with last principal factors (6 month interest rate)



The conditional marginal distribution for the 6 month US interest rate changes is generated with a Factor Quantile model based on the principal component representation in Equation 4.5 and the last five or four principal components. We use data from 03 January 1994 to 29 June 2018 and the quantile partition of Equation 4.4 for the calibration as illustrated with the rugs on the right-side axis.

first few principal components for forecasting applications. However, both regressions may provide an interesting view on the certainty of the distribution:

(i) The regression on the first few principal components uses a statistical factor model that describes the dependent variables accurately, especially if many principal components are considered. Therefore, a range outlined by a standard deviation multiple may be taken as the confidence interval within which we expect the true distribution to be. This may be applied in risk assessment settings where accurate forecasts are of secondary importance to certainty statements.

(ii) Furthermore, this range allows us to identify parts of the distribution where the uncertainty is particularly large. For instance, the distribution with the last four principal factors in Figure 4.6 is highly confident in the tails and centre estimation but less so in the areas in-between.

4.3.2 Bagging Quantile Forecasts

The discussion at the start of Section 4.3.1 revealed that a principal component representation using the first few components attributes too large a variance around the expectation to be directly useful for forecasting purposes. Rather than relying solely on the expectation of the future conditional quantiles, we now consider statistical techniques to extend our analysis to their entire distribution.

Focus for now on the i-th element of $\mathbf{y}_t$ with conditional quantiles $\mathbf{y}_t^{(\tau)}$, $\tau \in \mathbb{Q}$. To obtain the distribution of the conditional quantiles, we could generate m-variate draws from the empirical distribution of the principal components shown in Figure 4.7 and translate these to a distribution of the conditional quantiles. However, this method has several issues:

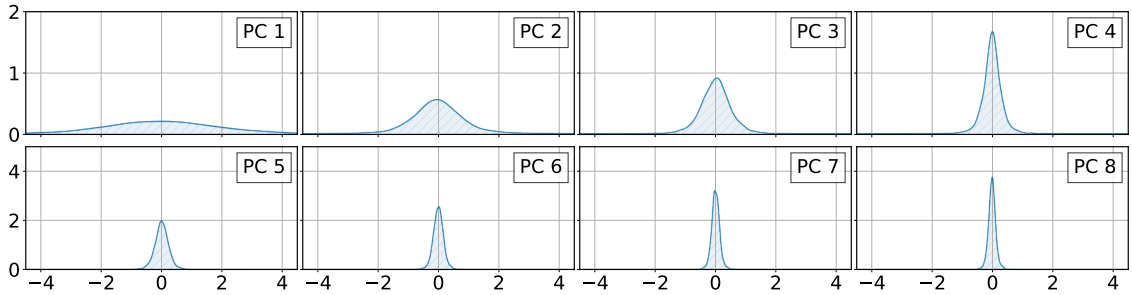
(i) Getting an appropriate sample that considers the dependency structure of the conditional quantiles is difficult. The principal components are uncorrelated but not independent and therefore we would need some additional restrictions on the simulations.

(ii) Sampling from the principal components may also be problematic. As statistical factors, principal components have no fundamental interpretation which complicates the choice of a parametric distribution. At the same time, non-parametric distributions could be inaccurate and add additional complexity to our methodology. In fact, if the principal component distributions were readily available, forecasting the distribution of $\mathbf{y}_{T+1}$ might become a moot enterprise altogether. Furthermore, the sampling size required to yield a good sampling distribution of the conditional quantiles increases drastically with m, further increasing the computational burden of this method.

An alternative approach is to utilize bootstrap aggregation or bagging by Breiman (1996) as a variance reduction technique. Suppose, training data

$$\mathbf{Z} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Figure 4.7: Principal component densities (US interest rate changes)



The principal components are based on the daily US interest rate changes used in Section 4.3.1, ranging from 03 January 1994 to 29 June 2018. We apply a Gaussian kernel to smooth the densities.

is used in combination with some procedure to obtain output $\hat{f}$. Then the meta-algorithm generates B bootstrap samples $\mathbf{Z}^1, \ldots, \mathbf{Z}^B$ of a pre-defined size by drawing from the original training data $\mathbf{Z}$ with replacement. The new bagging estimate is

$$\hat{f}^{\text{bag}} := \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b,$$

where $\hat{f}^b$ is the model estimate using $\mathbf{Z}^b$. Generally, $\hat{f}^{\text{bag}}$ has higher accuracy and lower variance than the original estimate $\hat{f}$. Bagging works particularly well if the procedure to estimate $\hat{f}$ from $\mathbf{Z}$ is unstable – see Hastie et al. (2009, pg. 282–288) for a detailed discussion.

In our case, we apply the asymptotic distribution of the sampling quantiles introduced in Section 3.1 to obtain bootstrap samples.[8] Given a quantile partition $\mathbb{Q} = (\tau_1, \ldots, \tau_q)$, Koenker and Bassett Jr (1978) show that the sample quantiles based on n observations

$$\hat{y}_{it}^{(\tau_1)}, \ldots, \hat{y}_{it}^{(\tau_q)}$$

are asymptotically normally distributed, that is

$$\sqrt{n} \left(\left(\hat{y}_{it}^{(\tau_1)}, \ldots, \hat{y}_{it}^{(\tau_q)} \right)' - \left(y_{it}^{(\tau_1)}, \ldots, y_{it}^{(\tau_q)} \right)' \right) \overset{n}{\rightsquigarrow} \mathcal{N}(\mathbf{0}, \mathbf{\Omega}). \tag{4.9}$$

The covariance is given by a matrix with elements

[8] For an easier discussion of the bagging algorithm, we focus on the conditional densities rather than the conditional distributions. In our continous case, a well-defined density function exists for each distribution function.

$$(\mathbf{\Omega})_{kl} = \frac{\tau_k(1-\tau_l)}{f\left(y_{it}^{(\tau_k)}\right) f\left(y_{it}^{(\tau_l)}\right)}, \tag{4.10}$$

where f is the density corresponding to the distribution of y_{it}. While the asymptotic mean and covariance are not available through Equation 4.10 because the density f is unavailable, we do know these values for $\hat{y}_{it}^{(\tau_1)}, \ldots, \hat{y}_{it}^{(\tau_q)}$ through Equations 4.6 and 4.8. Hence, the asymptotic distribution is known in case we apply the principal component representation.

We can use the bagging algorithm to reduce the variance of the distribution forecasts. For each draw of of the asymptotic distribution in Equation 4.9, we generate conditional density forecasts $\hat{f}_i^b|\hat{\mathbf{x}}_{T+1}$, $b = 1, \ldots, B$, and then combine them to

$$\hat{f}_i^{\text{bag}}|\hat{\mathbf{x}}_{T+1} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_i^b|\hat{\mathbf{x}}_{T+1}.$$

The bagging estimate is a well-defined density function, since it is non-negative and

$$\int \hat{f}_i^{\text{bag}}(y|\hat{\mathbf{x}}_{T+1})\,\mathrm{d}y = \int \frac{1}{B}\sum_{b=1}^{B} \hat{f}_i^b(y|\hat{\mathbf{x}}_{T+1})\,\mathrm{d}y = \frac{1}{B}\sum_{b=1}^{B}\int \hat{f}_i^b(y|\hat{\mathbf{x}}_{T+1})\,\mathrm{d}y = 1$$

due to the Fubini–Tonelli theorem. Thus, we can proceed as follows:

Stage 1 Calculate the first m principal components to derive the principal component representation of Equation 3.3;

Stage 2 Estimate quantile regressions for τ-quantiles where $\tau \in (0, 1)$ are prespecified by a partition $\mathbb{Q}$;

Stage 3 For each element i of the dependent variable, use the asymptotic normal distribution with expectation from Equation 4.6 and covariance from Equation 4.8 to sample associated conditional quantiles $\hat{y}_{it}^{(\tau_1)}, \ldots, \hat{y}_{it}^{(\tau_q)}$ and interpolate them to construct conditional marginal distributions;

Stage 4 Aggregate the sample conditional marginal distributions for each element i of the dependent variable to create the Factor Quantile conditional marginal distribution;

Stage 5 Use a conditional copula and apply Equation 4.2 to obtain the joint conditional distribution.

The aggregation in step 4 can, for instance, be done by sampling from each of the sample distributions and then combining all samples into a single distribution. Of course, there are several ways to generate the final distribution such as applying kernel density estimations or usinh the empirical distribution function. However, since the sample size can be chosen to be large, most methods should yield very similar results.

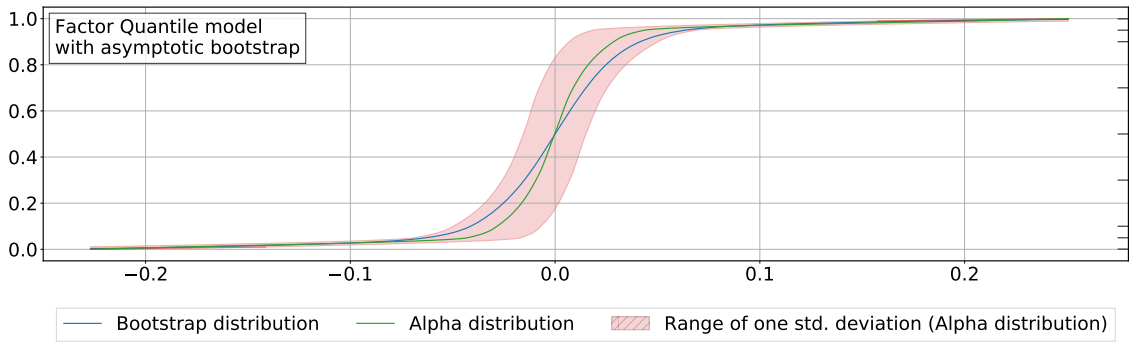
Figure 4.8 shows the resulting distribution forecast with 200 bagging replications in blue and the Alpha forecast as comparison in green. A red area illustrates the one-standard-deviation range from Figure 4.5. Both distributions are based on a principal component representation with the first principal component. Our asymptotic bagging distribution falls within the one-standard-deviation range but at the same time is visibly different from the Alpha distribution. This is encouraging since

(i) The one-standard-deviation range covers a 68% confidence interval for the Factor Quantile bagging distribution, assuming asymptotic normality. Hence, we do not expect the true distribution to deviate from the area.

(ii) Simultaneously, the distribution forecast differs significantly from the Factor Quantile model with the first principal component. This is because the bagging algorithm utilizes the entire distribution of the quantiles rather than focusing only on the their expectation.

We further compare the distributions from the two Factor Quantile specifications in our empirical study in Chapter 6.

Figure 4.8: Forecast with asymptotic bagging (6 month interest rate)



The conditional marginal distribution for the 6 month US interest rate changes is generated with a Factor Quantile model based on the principal component representation in Equation 4.5 and the first principal component. Additionally, we apply a bagging aggregation algorithm outlined in Algorithm 2. We use data from 03 January 1994 to 29 June 2018 and the quantile partition of Equation 4.4 for the calibration as illustrated with the rugs on the right-side axis.

This bagging Factor Quantile method generates the quantile estimates through draws from a multivariate normal distribution. Therefore, crossings can theoretically occur, especially in case $\mathbb{Q}$ outlines a dense grid. During our empirical study, we do not encounter such cases. This is likely due to (i) our uncorrelated latent factors which are ordered according to their variance explained and (ii) our choice of quantile partition, paired with the fact that the normal distribution is not heavy-tailed.

Algorithm 2 summarizes the Factor Quantile bagging approach based on asymptotic normality in pseudo-code.

Algorithm 2: Factor Quantile Model with asymptotic bagging

Input : Quantile partition $\mathbb{Q}$ with $0 < \tau < 1$ for all $\tau \in \mathbb{Q}$;
Observations on $\mathbf{y}_t$ for $t = 1, \ldots, T$;

Output : Conditional multivariate distribution $\hat{F}|\hat{\mathbf{x}}_{T+1}$ of $\mathbf{y}_t$;

Use observations to calculate the first $m \leq d$ principal components $\mathbf{x}_t = (p_{1t}, \ldots, p_{mt})$ where m is determined by the target for the variance explained;

for $i = 1, \ldots, d$ **do**

Estimate the factor quantile regressions

$$\mathrm{y}_{it}^{(\tau)} \leftarrow \alpha_i^{(\tau)} + \boldsymbol{\beta}_i^{(\tau)} \mathbf{x}_t + \varepsilon_{it}^{(\tau)}$$

which yields $\hat{\alpha}_i^{(\tau)}$ and $\hat{\boldsymbol{\beta}}_i^{(\tau)}$ for each $\tau \in \mathbb{Q}$;

Compute mean and covariance matrix for the quantiles as

$$\hat{\boldsymbol{\mu}}_i \leftarrow \left(\hat{\alpha}_i^{(\tau)} : \tau \in \mathbb{Q} \right), \qquad \hat{\mathbf{V}}_i \leftarrow \left(\sum_{i=1}^{m} \hat{\beta}_i^{(\tau_k)} \hat{\beta}_i^{(\tau_l)} \lambda_i \right)_{kl}$$

for $b = 1, \ldots, B$ **do**

Draw one d-dimensional sample $\mathbf{q}_b \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{V}}_i\right)$;

Interpolate $\mathbf{q}_b$ through shape-preserving interpolation to a distribution $\hat{F}_i|\mathbf{q}_b$;

end

Sample from $\hat{F}_i|\mathbf{q}_1, \ldots, \hat{F}_i|\mathbf{q}_B$ and aggregate samples to an estimate of $\hat{F}_i|\mathbf{x}_{T+1}$, the conditional distribution function of $\mathrm{y}_{i,T+1}$ with an empirical distribution function;

end

Generate the conditional multivariate distribution with the marginal distributions and a conditional copula

$$\hat{F}(\mathbf{y}|\hat{\mathbf{x}}_{T+1}) \leftarrow C\left(\hat{F}_1(y_1), \ldots, \hat{F}_d(y_d) \middle| \hat{\mathbf{x}}_{T+1} \right);$$

Chapter Five

Distribution Forecast Evaluation

5.1 Formal Tests of Forecast Performance 86
5.1.1 Diebold-Mariano Test 88
5.1.2 Model Confidence Sets 90
5.2 Proper Scoring Rules 93
5.2.1 Continuous Ranked Probability Scores 97
5.2.2 Energy Scores 101
5.2.3 Variogram Scores 103

The literature on forecast evaluation has evolved much in the past decades as is evident with the rejection of the now seminal paper Diebold and Mariano (1995) in 1991 motivated by the journal reviewer's disbelief that formal forecast evaluation is a necessary topic altogether (Diebold, 2015). With the increasing popularity of distribution forecasting, the need for techniques to measure the precision of probabilistic predictions rises accordingly (Elliott and Timmermann, 2008). Most of the traditional literature has focused on point predictions (Timmermann, 2000), although de Finetti (1975), Dawid (1984) and others have argued early on the importance of the probabilistic nature for forecasts. In this chapter we discuss various methods to compare and to quantify the accuracy of probabilistic forecasts that we apply in Chapters 6 and 7.

We follow Gneiting et al. (2007) and contend that the goal of distributional forecasting is to maximize the sharpness subject to calibration since sufficiently strong calibration conditions imply asymptotic equivalence to the ideal forecast. Calibration refers to the statistical consistency between a distributional forecast and the observations while sharpness refers to the concentration of a forecasted distribution, measured by the width of prediction intervals. As such, the two concepts are similar to unbiasedness and efficiency of statistical estimators. Heuristically, realisations should be indistinguishable from random draws of a calibrated forecast distribution (Gneiting and Katzfuss, 2014).

Definition 5.1 (Calibration)**.** At time $t = 1, \ldots, T$, let G_t and F_t be the continuous true distribution and forecast distribution respectively. $\{F_t\}_{t=1}^T$ is probabilistically calibrated relative to $\{G_t\}_{t=1}^T$ if

$$\frac{1}{T}\sum_{t=1}^{T} G_t \circ F_t^{-1}(p) \xrightarrow{T} p \quad \text{a.s.} \quad \forall p \in (0,1),$$

exceedance calibrated relative to $\{G_t\}_t$ if

$$\frac{1}{T}\sum_{t=1}^{T} G_t^{-1} \circ F_t(x) \xrightarrow{T} x \quad \text{a.s.} \quad \forall x \in \mathrm{dom}(F_t),$$

and marginally calibrated relative to $\{G_t\}_t$ if limits

$$\overline{G}(x) = \lim_{T\to\infty}\left\{\frac{1}{T}\sum_{t=1}^{T} G_t(x)\right\} \quad \text{and} \quad \overline{F}(x) = \lim_{T\to\infty}\left\{\frac{1}{T}\sum_{t=1}^{T} F_t(x)\right\}$$

exist and equal each other for all x, and if the common limit distribution places all mass on finite values.

Gneiting and Ranjan (2013) show that probabilistic and marginal calibration are necessary conditions for ideal forecasts.

Probabilistic calibration can be tested through the probability integral transform (PIT) introduced by Rosenblatt (1952), Dawid (1984) and Diebold et al. (1998). Given a series of forecast $\{F_t\}_{t=1}^{T}$ and realisations $\{y_t\}_t$, the PIT value is defined as

$$p_t = F_t(y_t)$$

and is uniformly distributed if and only if the forecast $\{F_t\}_{t=1}^{T}$ is probabilistically calibrated. However, uniformity is only a necessary condition for the ideal forecast but not a sufficient one since PIT cannot distinguish biased forecasts in some cases (Hamill, 2001). Multivariate extensions apply PIT stepwise and are discussed further in Brockwell (2007). For marginal calibration Gneiting et al. (2007) describe a test based on the similarity between the predictive and the empirical distribution but acknowledge that tests for exceedance calibration have not been designed.

There exist multiple alternatives to measure sharpness. For instance, sharpness for univariate distribution forecasts can be quantified by the range or variance of an ensemble. Various generalizations of this concept for multivariate probabilistic forecasts have been proposed, including the volume of the bounding box (Judd et al., 2007) or the root mean squared Euclidean distance between ensemble members and ensemble mean (Stephenson and Dolas-Reyes, 2000).

In our analysis, we quantify the calibration and sharpness through proper scoring rules and make inferences about the forecasting accuracy through formal tests of equal predictive ability.

5.1 Formal Tests of Forecast Performance

Following the paper by Diebold and Mariano (1995), many hypothesis testing procedures have been proposed to compare the accuracy of forecasts given a loss function. We introduce two prominent variations of these tests in the subsequent sections and use them in later chapters to formally evaluate the accuracy of (multivariate) distribution forecasts.

For simplicity, we limit the formal testing to the classic Diebold-Mariano (DM) test (Diebold and Mariano, 1995) and the more recent and easily interpretable Model Confidence Sets (MCS) by Hansen et al. (2011). There exist various extentions and alternatives of Diebold and Mariano (1995), such as the reality check of White (2000), the stepwise multiple testing procedure of Romano and Wolf (2005), the superior predictive ability test of Hansen (2005) and the conditional predictive ability test of Giacomini and White (2006):

(i) White (2000) compares a number of alternative forecasts to some benchmark forecast based on the null of equal forecasting performance but account for the effects of data snooping. Any superior performance that can be attributed to chance are neglected during the comparison. This framework is adjusted by Hansen (2005) who changes the test statistic and uses a sample-dependent distribution under the null hypothesis which increases the discrimination ability of the test.

(ii) Similary, Giacomini and White (2006) propose tests of equal predictive ability for the case when the forecasting models may be misspecified, allowing the data generating process to be heterogeneous. They account for estimation uncertainty and allow nested models.

(iii) Romano and Wolf (2005) introduce a stepwise comparison of multiple forecasts against a benchmark which applies multiple tests while controlling for type I errors. A joint confidence region created by bootstrap methods determines the hypotheses to reject at each step.

Our choice of the tests of equal performance is motivated by three reasons:

(i) All assumptions of DM and MCS are valid in our applications according to common testing procedures. The size of our studies also guarantees that the asymptotic inferences are accurate.

(ii) Additional features of alternative tests such as the consideration for model misspecifiation or data snooping are not necessary either because they are irrelevant for our model and data choices or because we already account for them beforehand – see Chapter 6 for additional details.

(iii) The results of DM and MCS complement each other since one compares two forecasts directly while the other ranks all forecasts through a sequential algorithm.

5.1.1 Diebold-Mariano Test

Traditionally, forecast evaluations have been assessed through Diebold-Mariano by testing the null hypothesis of no difference in the accuracy. The major advantages compared to simple statistics like the mean squared prediction error is that the forecast accuracy can be evaluated through a flexible loss function with weak assumptions about the forecast errors. These can be non-Gaussian, non-zero mean, as well as serially or contemporaneously correlated which is especially useful as competing forecasts often rely on overlapping information sets.

Let $\{\hat{y}_{1t}\}_{t=1}^{T}$ and $\{\hat{y}_{2t}\}_{t=1}^{T}$ be two forecasts for $\{y_t\}_{t=1}^{T}$ and let $\{L_{1t}\}_{t=1}^{T}$ and $\{L_{2t}\}_{t=1}^{T}$ be the corresponding losses for some arbitrary loss function L on the observation and the forecast. Diebold-Mariano focuses on the differences in the losses $d_t := L_{1t} - L_{2t}$ to test

$$\begin{aligned} \mathrm{H}_0 &: \mathbb{E}(d_t) = 0 \\ \mathrm{H}_A &: \mathbb{E}(d_t) \neq 0 \end{aligned}$$

through the test statistic

$$\mathrm{DM} = \frac{\overline{d}}{\sqrt{\frac{2\pi \hat{f}_{d_t}(0)}{T}}},$$

where

$$\overline{d} = \frac{1}{T} \sum_{i=1}^{T} (L_{1t} - L_{2t})$$

and $\hat{f}_{d_t}(0)$ is a consistent estimate of the spectral density of the loss differential at frequency 0

$$f_{d_t}(0) = \frac{1}{\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau), \qquad \gamma_{d_t}(\tau) = \mathbb{E}\left((d_t - \mathbb{E}(d_t))(d_{t-\tau} - \mathbb{E}(d_t))\right).$$

Given some parameter $S(T)$, known as the truncation lag, the denominator of the test statistic can be estimated through a weighted sum of sample covariances

$$2\pi \hat{f}_d(0) = \sum_{\tau=-(T-1)}^{(T-1)} \mathbb{I}\left(\frac{\tau}{S(T)}\right) \hat{\gamma}_d(\tau), \qquad \hat{\gamma}_d(\tau) = \frac{1}{T} \sum_{t=|\tau|+1}^{T} (d_t - \overline{d})(d_{t-|\tau|} - \overline{d}).$$

A popular choice for h-step-ahead forecasts is $S(T) = h-1$ since the optimal forecasts are at most $h-1$ dependent and therefore only $h-1$ sample autocovariances need to be considered.

Assuming the loss-differential series $\{d_t\}_{t=1}^{T}$ is covariance stationary and short memory, $\overline{d}$ is asymptotically normally distributed and

$$\text{DM} \sim \mathcal{N}(0,1).$$

To validate these assumptions, Diebold (2015) suggests tests for unit roots and other non-stationarities including trend, structural breaks or evolution as well as examinations of sample autocorrelation and spectrum. For finite samples, the use of the asymptotic normal distribution may not be warranted and can lead to wrong inferences by rejecting the null too often. Harvey et al. (1997) correct the bias for small sample sizes by introducing an adjusted Student-t distributed statistic as

$$\text{HLN} := \sqrt{\frac{T+1-2h+h(h-1)}{T}}\,\text{DM} \sim t_{T-1}.$$

5.1.2 Model Confidence Sets

Model confidence sets (MCS) by Hansen et al. (2011) rank the performance of competing models through sequential equivalence tests and an elimination rule based on a flexible loss function L, so that L_{it} is the loss of model i for a forecast at time t. The MCS algorithm takes an initial set of models $\mathcal{M}^0$ and returns $\mathcal{M}^*_{1-\alpha}$, an estimate of

$$\mathcal{M}^* := \left\{ i \in \mathcal{M}^0 : \mathbb{E}(d_{ij,t}) \leq 0 \text{ for all } j \in \mathcal{M}^0 \right\},$$

where $d_{ij,t} := L_{it} - L_{jt}$ as outlined in Algorithm 3. The performance of models in $\mathcal{M}^*_{1-\alpha}$ cannot be distinguished with equivalence tests at a confidence level of $1-\alpha$. Since the algorithm halts when the first hypothesis is accepted, MCS does not accumulate type I errors despite relying on sequential testing.

Algorithm 3: Model confidence set

1 Initiate the set of models as $\mathcal{M} = \mathcal{M}^0$;
2 Test $\mathrm{H}_0^{\mathcal{M}}$ at level α using the test statistic specified in Equation 5.1;
3 If $\mathrm{H}_0^{\mathcal{M}}$ is accepted, return $\mathcal{M}^* = \mathcal{M}$; otherwise apply the elimination rule in Equation 5.2 and repeat steps 2 and 3 with $\mathcal{M} \backslash \{e\}$;

For the definition of the equivalence test $\delta_{\mathcal{M}}$, consider a finite set $\mathcal{M}$ with models indexed by $i = 1, \ldots, N$. Then for $i, j = 1, \ldots, N$ and $t = 1, \ldots, T$ we assume:

(i) For some $r > 2$ and $\gamma > 0$, it holds that $\mathbb{E}(|d_{ij,t}|^{r+\gamma}) < \infty$ for all $i, j \in \mathcal{M}$;

(ii) $\{d_{ij,t}\}$ is strictly stationary and a mixing sequence with α of size $-r/(r-2)$ for all $i, j \in \mathcal{M}$.[1]

The hypothesis of the equivalence test are then set as

$$\mathrm{H}_0^{\mathcal{M}} : \mathbb{E}(d_{ij,t}) = 0 \text{ for all } i, j,$$
$$\mathrm{H}_A^{\mathcal{M}} : \mathbb{E}(d_{ij,t}) \neq 0 \text{ for some } i \neq j.$$

Similar to Diebold and Mariano (1995), we assess the null and its alternative by the test statistic

[1] We refer to Bradley (2005) for the definitions of the mixing conditions and the corresponding measures of dependence.

$$T_{\mathcal{M}} := \max_{i,j \in \mathcal{M}} \left| \frac{\overline{d_{ij}}}{\hat{\sigma}} \right|, \tag{5.1}$$

where

$$\overline{d_{ij}} = \frac{1}{T} \sum_{t=1}^{T} d_{ij,t}$$

is the average relative sample loss and $\hat{\sigma}^2$ is the bootstrapped estimate of the variance of $\overline{d_{ij}}$. The asymptotic distributions of $T_{\mathcal{M}}$ are non-standard and are therefore estimated through a bootstrap procedure. This also avoids high-dimensional covariance matrices which can be computationally intensive and challenging (White, 2000).

An elimination rule $e_{\mathcal{M}}$ identifies the worst model e if the hypothesis of equal predictive ability is rejected.[2] The worst model

$$e = \arg\max_{i \in \mathcal{M}} \left\{ \sup_{j \in \mathcal{M}} \frac{\overline{d_{ij}}}{\hat{\sigma}} \right\}, \tag{5.2}$$

is the one for which exclusion may lead to a reduction in the test statistic.

The set $\mathcal{M}^*_{1-\alpha}$ includes the best models of $\mathcal{M}^0$ with a certain probability. Similar to the concept of confidence intervals, the number of models in the MCS increases as we decrease α. Hansen et al. (2011) show that the sequential testing procedure guarantees

$$\lim_{T \to \infty} P(\mathcal{M}^* \subset \mathcal{M}^*_{1-\alpha}) \geq 1 - \alpha,$$

where $\mathcal{M}^*$ is the unknown set that contains the best models with respect to the loss function L.

MCS p-values offer an easy way to discern whether a model is included in a certain model confidence set $\mathcal{M}^*_{1-\alpha}$. Suppose the MCS algorithm terminates after k iterations. Denote the sets created by the elimination rule by $\mathcal{M}^0 \supset \mathcal{M}_1 \supset \ldots \supset \mathcal{M}_k$

[2] See Hansen et al. (2011) for a general discussion on the requirements of valid test and elimination rule combinations.

and the p-value of model i corresponding to the equivalence test on the set $\mathcal{M}_l$ by $p_{i,\mathcal{M}_l}$ with the convention that $p_{i,\mathcal{M}_k} \equiv 1$. Then, the MCS p-value of model i is defined as

$$p_i := \max_{l \leq k} p_{i,\mathcal{M}_l}$$

and model i is included in $\mathcal{M}^*_{1-\alpha}$ if and only if $p_i \geq \alpha$.

This sequential testing methodology offers multiple advantages compared to classic hypothesis tests of equal forecasting accuracy:

(i) First, MCS considers that in many applications the data is not informative enough to select a best model unequivocally or a single dominating model does not exist. Therefore, the superior set of models $\mathcal{M}^*_{1-\alpha}$ may contain multiple models which cannot be distinguished at a certain confidence level $1-\alpha$ and assigns each model with a significance value p.[3] In contrast, classic hypothesis tests such as Diebold-Mariano can only compare models pairwise, leading to $N(N-1)/2$ separate tests which are more difficult to interpret than a superior set $\mathcal{M}^*_{1-\alpha}$ and which might accumulate the type I errors (Leeb and Pötscher, 2003);

(ii) Second, the methodology allows for arbitrary loss functions. This enables the flexible application of user-specified criteria which might be more adapted than standard loss functions such as the symmetric mean square prediction error;

(iii) Third, there is no need for any benchmark models in contrast to other evaluation methodologies, such as the reality check for data snooping (White, 2000) or the test for superior predictive ability (Hansen, 2005). Hence, MCS can be used in model selection applications.

[3]This, conversely means that not all models in the MCS may be good models. Only models which are significantly inferior to other models in the initial set $\mathcal{M}^0$ are eliminated during the sequential testing.

5.2 Proper Scoring Rules

Forecasting accuracy evaluations introduced in Section 5.1 rely on loss measures to quantify the performance of a distribution forecast. As Diebold and Mariano (1995) mention, this loss generally depends on the underlying economic structures associated with the forecast which means simple statistical measures such as the mean squared prediction error (MSPE) are often inadequate. Scoring rules offer a promising alternative by condensing the accuracy of a distribution forecast to a single penalty oriented value while retaining attractive statistical properties.

Definition 5.2 (Scoring rule)**.** Let $\mathcal{F}$ be the convex class of distributions on $(\Omega, \mathcal{A})$. A scoring rule is a function

$$S : \mathcal{F} \times \Omega \longrightarrow \mathbb{R} \cup \{-\infty, \infty\}$$

that assigns each distribution of $\mathcal{F}$ a certain score.

A scoring rule S is proper if and only if for all distributions F and G with associated densities f and g

$$\mathbb{E}_F S(F, Y) = \int f(y) S(F, y) \, \mathrm{d}y \leq \int f(y) S(G, y) \, \mathrm{d}y = \mathbb{E}_F S(G, Y). \tag{5.3}$$

Further, a scoring rule is strictly proper if Equation 5.3 holds with equity only for $F = G$ almost surely.

Propriety of a scoring rule is a necessary condition since the ideal forecast is preferred irrespective of the cost-loss structure (Diebold et al., 1998; Granger and Pesaran, 2000). A proper scoring rule is designed so that a forecaster who believes the future distribution to be F has no incentive to predict any distribution $G \neq F$ (Gneiting et al., 2007). The term has been coined by Winkler (1996, 1977) who shows that proper scoring rules test for both calibration and sharpness of a distribution forecast simultaneously. The usage of non-proper scoring rules is generally not recommended since those can lead to wrong inferences (Gneiting and Ranjan, 2011).

Despite the focus of the literature on propriety, it is important to note that this property by itself is not a sufficient condition for a good accuracy measure. Every constant scoring rule is by definition proper but obviously useless for forecast evaluation. Even strictly proper scoring rules which have to assign non-constant values to distinguish the ideal distribution can be problematic since the comparison in Equation 5.3 is between the true distribution F and some forecasted distribution G. Since F is generally unknown, applications of scoring rules compare predictions G and G' which both receive a higher score than F by a strictly proper scoring rule but there is no guarantee that the preferred distribution receives the lower score. We discuss this point further in our simulation study in Chapter 7.

Scoring rules can be used to measure the forecasting accuracy of both univariate and multivariate distribution forecasts. However, since the degrees of freedom increase rapidly in higher dimensions, the encapsulation into a single score is associated with a loss of information. Most notably, multivariate scores tend to focus on the dependency structure, neglecting individual marginal performances. Therefore, various multivariate and univariate scores compliment each other and should be employed in combination during higher-dimensional distribution evaluation as recommended by both Gneiting et al. (2008) and Scheuerer and Hamill (2015).

To focus on a clear message, we confine our discussion in the subsequent sections to the continuous ranked probability score (CRPS) as well as the energy score and variogram score. There are multiple popular (strictly) proper univariate and multivariate alternatives including the logarithmic, the quadratic and the pseudo-spherical score as well as the Dawid-Sebastiani score (Gneiting and Ranjan, 2011; Scheuerer and Hamill, 2015). For those, we provide a definition and briefly explain the reasoning behind their exclusion.

Definition 5.3 (Logarithmic, quadratic and pseudo-spherical score)**.** Let y be an observation of the random variable Y and let F be a forecast of the distribution of Y with density f. Further, let μ be a σ-finite measure on the measurable space $(\Omega, \mathcal{A})$ and define

$$\|f\|_\alpha := \left(\int f(y)^\alpha \mu \, \mathrm{d}y \right)^{1/\alpha}.$$

Then the logarithmic, quadratic and pseudo-spherical scores are defined as

$$\begin{aligned} \mathrm{LogS}(F, y) &= -\log(f(y)), \\ \mathrm{QS}(F, y) &= 2f(y) - \|f\|_2^2, \\ \mathrm{PseudoS}(F, y) &= f(y)^{\alpha-1} / \|f\|_\alpha^{\alpha-1}. \end{aligned}$$

The spherical score is a special case of the pseudo-spherical score with $\alpha = 2$. All three scores are strictly proper under certain conditions but we prefer CRPS to them:

(i) The alternative univariate scores rely on predictive densities which might not be available, especially with ensemble forecasts;

(ii) Additionally, they only credit forecasts for high probabilities of the realizing value but not for high probabilities to values near the realizing one (Gneiting and Raftery, 2007).

The Dawid-Sebastiani score by Dawid and Sebastiani (1999) is a multivariate score that depends solely on the mean and covariance of the forecasts. It is proper relative to the class of distributions with finite second moments and strictly proper if additionally the distributions are fully characterized by the first two moments.

Definition 5.4 (Dawid-Sebastiani score)**.** Let $\mathbf{y} = (y_1, \ldots, y_d)'$ be an observation of the random vector $\mathbf{Y}$ and let F be a forecast of the distribution of $\mathbf{Y}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then the Dawid-Sebastiani score is defined as

$$\mathrm{DS}(F, \mathbf{y}) = \log(\det \boldsymbol{\Sigma}) + (\mathbf{y} - \boldsymbol{\mu})' \, \boldsymbol{\Sigma}^{-1} \, (\mathbf{y} - \boldsymbol{\mu}).$$

We exclude this score in our multivariate evaluation despite the advantages of including multiple multivariate measures for two reasons:

(i) The score only relies on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and cannot distinguish forecasts with differences only in higher moments which is often detrimental in financial and economic data sets;

(ii) Also, accurate estimation estimation of the covariance matrix $\boldsymbol{\Sigma}$ is a challenging task (White, 2000).

5.2.1 Continuous Ranked Probability Scores

The continuous ranked probability score (CRPS) introduced by Matheson and Winkler (1976) and augmented by Gneiting and Ranjan (2011) is a widely used generalization of the mean absolute error and properly compares distribution forecasts with a potential focus on certain regions of interest.

Definition 5.5 (CRPS)**.** Let y be an observation of the random variable Y and let F be a forecast of the distribution of Y with density f. Then, the continuous ranked probability score is defined as

$$\mathrm{CRPS}_\nu(F, y) = \int_0^1 \mathrm{QS}_\alpha(F^{-1}(\alpha), y)\nu(\alpha)\,\mathrm{d}\alpha,$$

where $\nu : [0, 1] \to \mathbb{R}_{\geq 0}$ is a quantile weight function and QS_α is the quantile score

$$\mathrm{QS}_\alpha(F^{-1}(\alpha), y) = 2(\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha)(F^{-1}(\alpha) - y).$$

Apart from the quantile score representation in Definition 5.5, CRPS can also be expressed using the Brier probability score through

$$\begin{aligned} \mathrm{CRPS}_u(F, y) &= \int_{-\infty}^{\infty} \mathrm{PS}(F(z), \mathbb{1}\{y \leq z\})u(z)\,\mathrm{d}z, \\ \mathrm{PS}(F(z), \mathbb{1}\{y \leq z\}) &= (F(z) - \mathbb{1}\{y \leq z\})^2 \end{aligned} \tag{5.4}$$

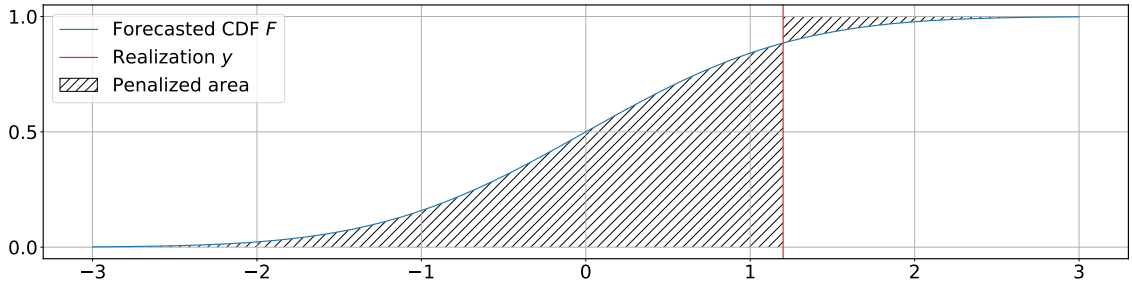
with threshold weight function $u : \mathbb{R} \to \mathbb{R}_{\geq 0}$ as shown by Laio and Tamea (2007). Given a realization y, the integral of Equation 5.4 splits into two easily interpretable parts which get penalized by the score as visualized in Figure 5.1. Furthermore, it follows directly that CRPS is equivalent to the mean absolute error for any deterministic forecast.

Additionally, Gneiting and Raftery (2007) derive the kernel score representation

$$\mathrm{CRPS}_u(F, y) = \mathbb{E}_F\left(Y - y\right) - \frac{1}{2}\mathbb{E}_F\left(Y - Y'\right), \tag{5.5}$$

where Y and Y' are independent random variables with sampling distribution F. This concise expression serves as a foundation for the generalization of CRPS to the multivariate energy score discussed in Section 5.2.2.

Figure 5.1: CRPS Schematic



We use $F = \Phi$ and $y = 1.2$ to illustrate the concept of CRPS. The forecasted distribution F is penalized for the shaded area left and right of the realized value y through $\int_{-\infty}^{y} F(z)^2 \, \mathrm{d}z$ and $\int_{y}^{\infty} (1 - F(z))^2 \, \mathrm{d}z$ respectively. A low score suggests high sharpness of the distribution forecast around the realisation. Here, φ and Φ denote the density and the distribution of the standard normal distribution.

For densities with finite first moment, CRPS is strictly proper. Densities with infinite first moments in contrast have infinite CRPS. Thus, the true probability function receives the lowest CRPS and is preferred to any other probabilistic forecast. Compared to other univariate proper scores such as the logarithmic, quadratic or (pseudo-)spherical score, CRPS does not harshly penalize unlikely events and is thus less sensitive to outliers (Selten, 1998).

Emphasizing specific parts of the distribution by the choice of the quantile or threshold weight functions is simple since any non-negative function can be used. If the threshold weight function is integrable, the corresponding CRPS is finite and bounded by the integral of the weight function. Table 5.1 lists the proposed functions by Amisano and Giacomini (2007) that we use in Chapter 6.

Computations of CRPS are generally efficient since closed-form expressions for many common distributions are available.[4] In case F is an empirical distribution function, the integral in Equation 5.4 breaks down to discrete finite sums and can be calculated with computational complexity $\mathcal{O}(n \log n)$ as described by Grimit et al. (2006).

[4]See Jordan et al. (2017) for an overview of distributions with closed-form expressions.

Table 5.1: Possible weights for CRPS

Emphasis	Quantile weights	Threshold weights
Uniform	$\nu(\alpha) = 1$	$u(z) = 1$
Centre	$\nu(\alpha) = \alpha(1-\alpha)$	$u(z) = \varphi(z)$
Both tails	$\nu(\alpha) = (2\alpha-1)^2$	$u(z) = 1 - \varphi(z)/\varphi(0)$
Right tail	$\nu(\alpha) = \alpha^2$	$u(z) = \Phi(z)$
Left tail	$\nu(\alpha) = (1-\alpha)^2$	$u(z) = 1 - \Phi(z)$

The weight functions $\nu : [0,1] \to \mathbb{R}_{\geq 0}$ and $u : \mathbb{R} \to \mathbb{R}_{\geq 0}$ put additional emphasis on certain parts of the distribution. Forecasts which deviate on those parts are penalized additionally and receive a higher CRPS. Here, φ and Φ denote the density and the distribution of the standard normal distribution.

Comparisons between different forecasts can be made through their average scores either directly as an omnibus performance measure or through hypothesis tests. Let $\{\mathbf{Z}_t\}_{t=1}^T$ be a stochastic process that can be partitioned as $\mathbf{Z}_t = (Y_t, \mathbf{X}_t)$ where Y_t is the forecasted variable and $\mathbf{X}_t$ is a vector of predictors. Furthermore suppose $T = m+n$. Denote by f_{t+k} and g_{t+k} two density forecasts for Y_{t+k} that are generated for $t = m, \ldots, m+n-k$ and which depend only on $\mathbf{Z}_{t-m+1}, \ldots, \mathbf{Z}_t$. Given the average scores

$$\overline{\mathrm{CRPS}}_n^f = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{CRPS}(f_{t+k}, y_{t+k}),$$
$$\overline{\mathrm{CRPS}}_n^g = \frac{1}{n-k+1} \sum_{t=m}^{m+n-k} \mathrm{CRPS}(g_{t+k}, y_{t+k}),$$

the test of equal forecast performance is then based on

$$t_n = \sqrt{n} \frac{\overline{\mathrm{CRPS}}_n^f - \overline{\mathrm{CRPS}}_n^g}{\hat{\sigma}_n},$$

where

$$\hat{\sigma}_n^2 = \frac{1}{n-k+1} \sum_{j=-(k-1)}^{k-1} \sum_{t=m}^{m+n-k-|j|} \Delta_{tk} \Delta_{t+|j|,k},$$
$$\Delta_{tk} = \mathrm{CRPS}(f_{t+k}, y_{t+k}) - \mathrm{CRPS}(g_{t+k}, y_{t+k}).$$

The test statistic t_n is asymptotically standard normal under the null hypothesis of vanishing expected score differentials assuming:

(i) The weight function is bounded and non-negative;

(ii) $\hat{\sigma}^2$ is a consistent estimator of σ^2;

(iii) The moments

$$\int_{-\infty}^{\infty} f_{t+k}(x)|x|\,\mathrm{d}x, \qquad \int_{-\infty}^{\infty} g_{t+k}(x)|x|\,\mathrm{d}x, \qquad \mathbb{E}(|Y_{t+k}|),$$

are finite for all t. This condition is not necessary in case the threshold weight u is integrable;

(iv) $\{\mathbf{Z}_t\}_{t=1}^{T}$ is a mixing sequence with φ of size $-r/(2r-2)$, $r \geq 2$, or with α of size $-r/(r-2)$, $r > 2$.[5]

While the assumptions for the asymptotic distribution cannot be verified in practice, they should hold in most cases. Gneiting and Ranjan (2011) propose the rule of thumb that the asymptotic normality is appropriate unless the densities have infinite moments of low order.

[5]We refer to Bradley (2005) for the definitions of the mixing conditions and the corresponding measures of dependence.

5.2.2 Energy Scores

The energy score is a popular multivariate strictly proper score introduced by Gneiting and Raftery (2007) which generalizes the kernel representation of CRPS in Equation 5.5. It computes a weighted distance between the characteristic function of F and the characteristic function of the point measure at the value it realizes.

Definition 5.6 (Energy score)**.** Let $\mathbf{y} = (y_1, \ldots, y_d)'$ be an observation of the random vector $\mathbf{Y}$ and let F be a forecast of the distribution of $\mathbf{Y}$ such that $\mathbb{E}_F(\|\mathbf{Y}\|^\beta)$ is finite. The energy score is then defined as

$$\mathrm{ES}_\beta(F, \mathbf{y}) = \frac{1}{2}\mathbb{E}_F\left(\|\mathbf{X} - \mathbf{X}'\|^\beta\right) - \mathbb{E}_F\left(\|\mathbf{X} - \mathbf{x}\|^\beta\right),$$

where $\mathbf{X}$ and $\mathbf{X}'$ are independent random vectors with distribution F.

Székely (2003) shows that the energy score with $\beta \in (0, 2)$ is strictly proper while Gneiting and Raftery (2007) provide an alternative and more general proof. In case $\beta = 2$, the energy score is proper but not strictly proper since it reduces to the squared error

$$\mathrm{ES}_2(F, \mathbf{y}) = -\|\boldsymbol{\mu}_F - \mathbf{y}\|^2,$$

where $\boldsymbol{\mu}_F$ is the mean vector associated with F.

In practice, usually $\beta = 1$ as the energy score reduces to the CRPS in the univariate case for this parameterisation. Further, this yields a strictly proper score that is easier to compute than alternative values of β. If the components vary largely in magnitude, standardisations might be necessary.

Closed form solutions of the energy score are generally unavailable which means that computations are done through Monte Carlo methods. In case the prediction is provided in the form of an ensemble forecast of m discrete samples, the energy score for $\beta = 1$ reduces to

$$\widehat{\mathrm{ES}}_1(F, \mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{X}_i - \mathbf{y}\| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\mathbf{X}_i - \mathbf{X}_j\|. \tag{5.6}$$

Despite its popularity, this score has been criticized for being insensitive to misspecification of the dependency structure (Pinson and Girard, 2012; Pinson and Tastu, 2013) and for being unable to distinguish a good representation of the predictive distribution from a very sparse one (Scheuerer and Hamill, 2015).

5.2.3 Variogram Scores

An alternative multivariate score is the variogram score by Scheuerer and Hamill (2015) which is based on the concept of variograms from geostatistics. Similar to diagnostic methods by Hamill (2001) and Feldmann et al. (2015), the score examines pairwise element differences of the dependent variable $\mathbf{Y}$.

Definition 5.7 (Variogram score)**.** Let $\mathbf{y} = (y_1, \ldots, y_d)'$ be an observation of the random vector $\mathbf{Y}$ and let F be a forecast of the distribution of $\mathbf{Y}$ for which the p-th moment exists. Then the variogram score of order p is defined as

$$\mathrm{VS}_p(F, \mathbf{y}) = \sum_{i=1}^{d} \sum_{j=1}^{d} \left(|y_i - y_j|^p - \mathbb{E}_F \left(|X_i - X_j|^p \right) \right)^2 ,$$

where X_i and X_j are the i-th and j-th component of a random vector $\mathbf{X}$ with distribution F.

Scheuerer and Hamill (2015) show that the score is proper relative to the class of distributions for which the $2p$-th moments of all elements are finite. The variogram score is not strictly proper because they depend only on the p-th absolute moment of the distribution of the element differences. Therefore, it cannot distinguish any distributions where the element differences deviate in higher moments of order greater than p but are the same for moments of order less than or equal p.

Intuitively, the score makes use of the variogram of order p

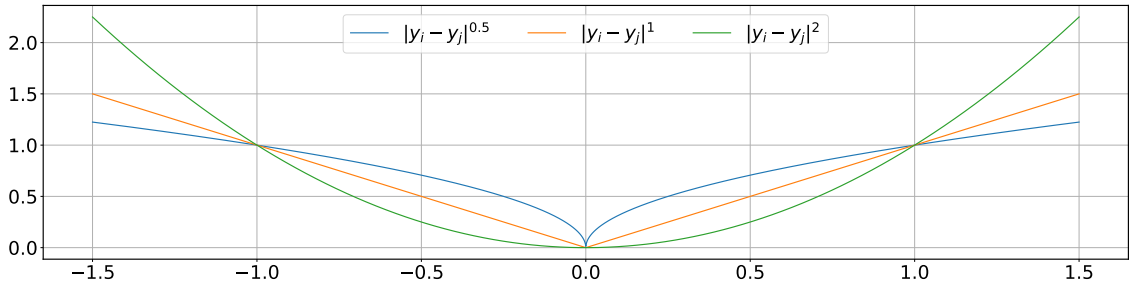
$$\gamma_p(i, j) = \frac{1}{2} \mathbb{E} \left(|X_i - X_j|^p \right) ,$$

which quantifies the degree of spatial dependence of a stochastic process. Pairwise comparisons measure the closeness of the deviations in the observations with those of the corresponding expectations.

The choice of p depends on the forecasted distribution and should generally be large enough to consider all relevant moments of the pairwise deviations but not too large to overly emphasize outliers through the exponentiation. Often, values $p = 0.5, 1, 2$ are suggested which are based on the rodogram ($p = 0.5$), mandogram

($p = 1$) and variogram ($p = 2$) respectively. Figure 5.2 shows the effect of p by illustrating the observed variogram $|y_i - y_j|^p$ of different popular orders relative to changes in $|y_i - y_j|$. It is clearly visible that the magnitude of the effect depends heavily on the value of $|y_i - y_j|$ with the absolute slope varying between 0 and 3 in the depicted domain $(-1.5, 1.5)$. The sensitivity of the observed variogram in a neighbourhood of zero deviation is strongest for $p = 0.5$ and very weak for $p = 2$. This order reverses for $|y_i - y_j| > (1/4)^{2/3}$. As the acuteness of the variogram is an indicator of that of the variogram score, we expect parameter $p = 0.5$ to be more influential for similar y_i and y_j while $p = 2$ reacts more strongly when $|y_i - y_j|$ is expected to be large. The choice $p = 1$ is able to differentiate in both cases.

Figure 5.2: Variogram observation of various orders



The figure shows the effect of the variogram order depending on the observed absolute difference $|y_i - y_j|$. Slight deviations in $|y_i - y_j|$ affect the observed variogram $|y_i - y_j|^p$ differently, depending on its order p.

As with the energy score, the encapsulation of the information to a single score leads to a loss of information. However, empirical applications support the sensitivity of the score to flawed forecasts, especially regarding the dependency structure (Scheuerer and Hamill, 2015).

Since the score is based on pairwise deviations, any bias that is the same for all components of the forecast cancel out and are therefore undetectable. This further motivates the practice to use multiple proper scores for the evaluation of multivariate distribution forecasts.

Approximations of the variogram score given an ensemble forecast are easy to calculate through

$$\widehat{\mathrm{VS}}_p(F, \mathbf{y}) = \sum_{i=1}^{d} \sum_{j=1}^{d} \left(|y_i - y_j|^p - \frac{1}{n} \sum_{k=1}^{n} \left| X_i^{(k)} - X_j^{(k)} \right|^p \right)^2 ,$$

where $X_i^{(k)}$ is the i-th element of sample k of the ensemble forecast.

Chapter Six

Empirical Study

6.1 Data Description 110
6.2 Empirical Design 117
6.3 Forecasting Accuracy Results 124
6.3.1 Univariate Forecasting Accuracy 126
6.3.2 Multivariate Forecasting Accuracy 140

In this chapter, we compare the out-of-sample performance of our Factor Quantile models from Chapter 4 through the evaluation methods described in Chapter 5.

Previous studies of forecasting models with quantile regression usually only include a limited empirical evaluation or, like Ma and Pohlman (2008) or Zhu (2013), exclude them entirely. Common issues include:

Short out-of-sample periods: Many studies use short out-of-sample periods. For instance, the empirical evaluation of Koenker and Bassett (2010), Gaglianone and Lima (2012) and Manzan (2015) are based on only 48, 77 and 438 observations respectively. This may not be sufficient to yield general results for the relative forecasting accuracy of the distribution forecasting methods.

Weak benchmark models: Manzan (2015) and Meligkotsidou et al. (2019) use an autoregressive process that is encompassed by their quantile model as benchmark. The higher relative accuracy is therefore expected since the quantile model incorporates strictly more information than the benchmark and does not get penalized for the excess parameters during testing. Similarly, Cenesizoglu and Timmermann (2008) and Gaglianone and Lima (2012) apply symmetric GARCH models on data with monthly or quarterly frequency. These GARCH models cannot reflect the asymmetric properties of the data adequately and may be unsuited as benchmark for these low frequencies because volatility clustering is typically only present in data with daily or higher frequency. Furthermore, Gaglianone and Lima (2012) calibrate the GARCH models on 90 to 166 observations. This may be insufficient to accurately estimate the GARCH parameters or the long-term volatility.

Improper evaluation: Most studies do not use proper scoring rules and limit their evaluation to simple statistics such as the conditional coverage tests of Christoffersen (1998) or simplified statistical measures such as the root mean square error (RMSE) or median absolute deviation (MAD). This is for instance the case for Gaglianone and Lima (2012), Pedersen (2015) and Bunn et al.

(2016). Only few studies such as Manzan (2015) and Meligkotsidou et al. (2019) apply proper quantile scores, but even then the results are difficult to interpret. Manzan (2015) examines several quantiles separately which leads to 468 test statistics. The large amount of tests accumulates type I errors and further complicates the identification of the most accurate distribution forecast because the best model varies across the quantiles.

We assess the forecasting accuracy of our Factor Quantile methodology rigorously in an extensive empirical study with two standard econometric model classes for forecasting systems of exchange rates, the term structure of interest rates and commodity future indices. As discussed in Chapter 2, similar data sets have been used in the forecasting literature, for instance by Greenaway-McGrevy et al. (2018) for US exchange rates, by Zolotko and Okhrin (2014) and Chen et al. (2014) for commodities and by Almeida et al. (2017) for the US treasury yield curve. We quantify the accuracy of all distribution forecasts using univariate and multivariate proper scoring rules as well as Model Confidence Sets (MCS) which avoids large numbers of test statistics by ranking the performance of all models directly. Combined, we have an out-of-sample period that includes over 12,000 observations that we examine over the entire sample period as well as over sub-periods.

Section 6.1 begins with a description of the three data sets we use and points out striking features. We restrict all our multivariate systems to eight assets as some benchmark models struggle with the application in even higher dimensions. The choice of the data is especially important because data snooping effects will affect the results in case the assets are not properly motivated but picked selectively. Utilizing the theoretical background of Chapter 5, we detail the methodology of our forecasting accuracy evaluation that uses the continuous ranked probability score (CRPS), the energy score and the variogram score to properly quantify the performance in Section 6.2. The MCS ranks each model based on its respective scores. Section 6.3 presents our results for the entire sample and for specific sub-samples. Many results cannot be reported in detail for reasons of space, but they are available

as supplementary materials electronically, along with the data and code used to generate these results.

6.1 Data Description

Our empirical study involves eight-dimensional time series on USD-denominated exchange rates, US interest rates and Bloomberg investable commodity indices of daily frequency. Through these diverse data sets, we illustrate the performance of Factor Quantile models relative to the benchmark models in different applications and establish our semi-parametric model as a general methodology. We obtain the daily exchange rates and commodity index values from Thomson Reuters Datastream and the interest rates data from the US Treasury website. All time series end on 30 June 2018 but the start date varies with data availability. Within each set we have selected variables to broadly represent the asset class:[1]

Exchange rate returns: The exchange rates are those with the highest trading volume excluding Chinese Renminbi, which was pegged to the USD until recently (Bank of International Settlements, 2016). Our data starts in January 1999 with the introduction of the Euro as accounting currency.

Interest rate changes: The interest rates span the term structure of US Treasury bonds from 6 months to 20 years. Alternative available maturities are 1 month, 2 month, 3 month, and 30 years but those miss data for an extended period of time and are therefore excluded. Our data starts in January 1994 after the 20-year maturity interest rate becomes available in October 1993.

Commodity index returns: The commodity indices are chosen to reflect the most liquid commodities with the highest USD-weighted production value and are diversified to represent the energy, grains, industrial / precious metals, softs and livestock sectors (Bloomberg, 2017). The Bloomberg commodity indices were launched in 1998 with a backward projection to January 1991. We include all available data in our study.

[1]Tables B1 and B2 contain the extracts from Bank of International Settlements (2016) and Bloomberg (2017) that motivate our choice of the assets within exchange rates and US interest rates.

We summarize the total sample period and the starting date of our out-of sample evaluation in Table 6.1.

Table 6.1: Sample for each data set

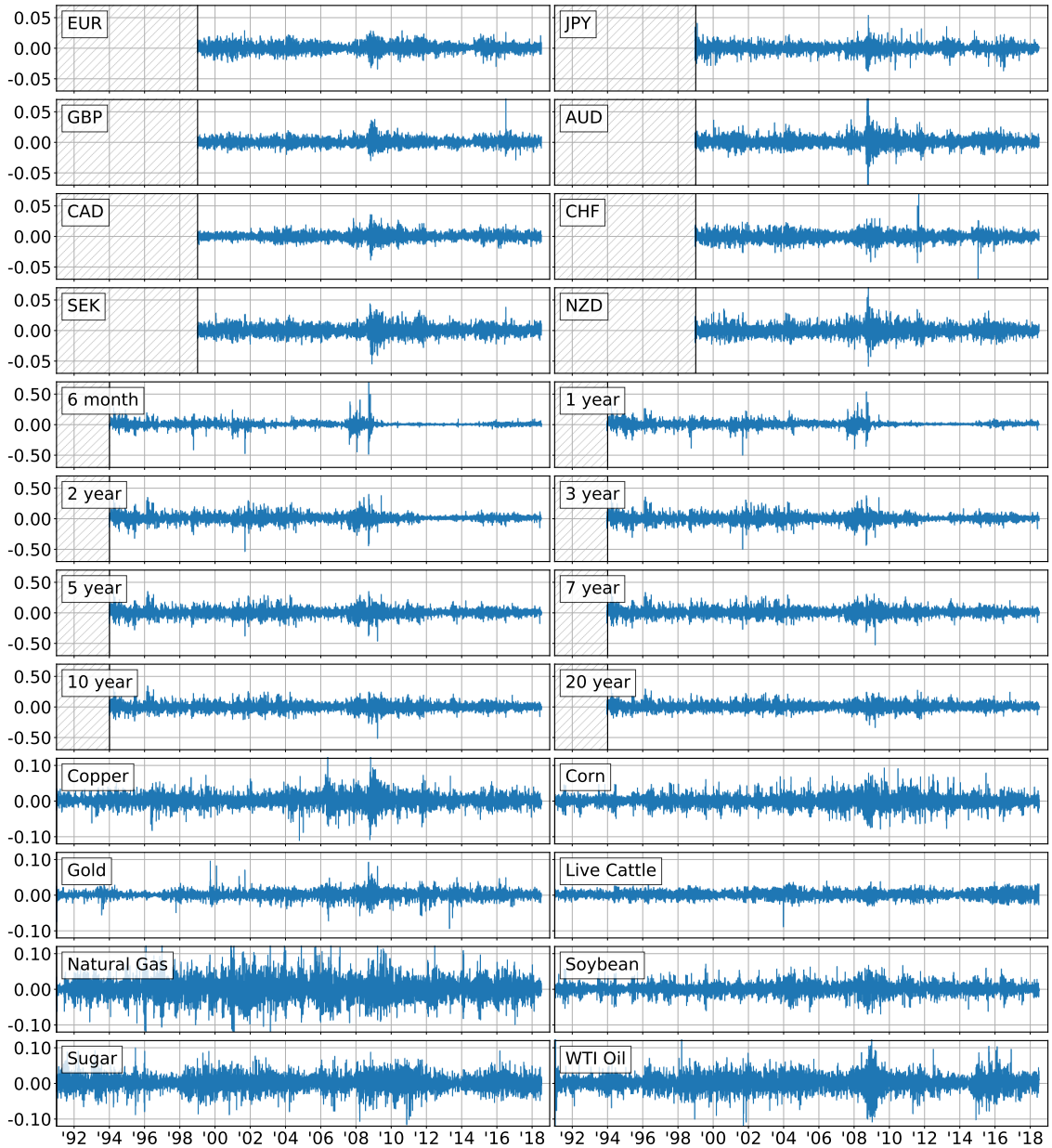
Data set	First date	Start evaluation	End date
Exchange rate returns	01 January 1999	28 February 2007	30 June 2018
Interest rate changes	01 January 1994	03 July 2002	30 June 2018
Commodity index returns	01 January 1991	26 February 1999	30 June 2018

All three data sets use daily frequencies, yielding over 18,000 observations in total. Out-of-sample evaluation starts after a calibration period which is discussed further in Section 6.2. The first dates vary due to data availability.

Figure 6.1 depicts the data employed, i.e. daily returns on exchange rates and commodity indices as well as daily basis-point changes in interest rates. This highlights the range, volatility and other idiosyncratic features of each system. For instance, US interest rates became highly volatile during the credit crunch of 2007, but were very stable during the last few years, particularly at the short end. Commodities have the greatest volatility overall, especially natural gas and sugar but their volatility has been generally increasing with the globalisation and accompanied financialization of commodity markets since 2003. Exchange rates are much less volatile than commodities, although a burst of volatility is evident soon after the banking crisis began in late 2008. The effects of the Brexit vote on the GBP in June 2016, and of the CHF devaluation in early 2015 are easily discernible. Summary statistics of the data are listed for monthly returns / changes in Table 6.2. The sample mean for all monthly returns and changes is around zero which allows us to apply the principal component representation of Factor Quantile models without prior transformations. Furthermore, all assets are leptokurtic and require heavy tailed distributions.

Considering the US interest rates term structure, note that the rates follow several different regimes, depicted in Figure 6.2. The term structures move between contango and backwardation, as well as through periods of growth and decline. A similar

Figure 6.1: Daily returns / changes on all three data sets



The exchange rate and commodity index data are obtained from Thomson Reuters Datastream and the US interest rates data are obtained from the US Treasury website. Each time series in the data sets includes 5,085, 7,173 and 6,130 daily realisations respectively.

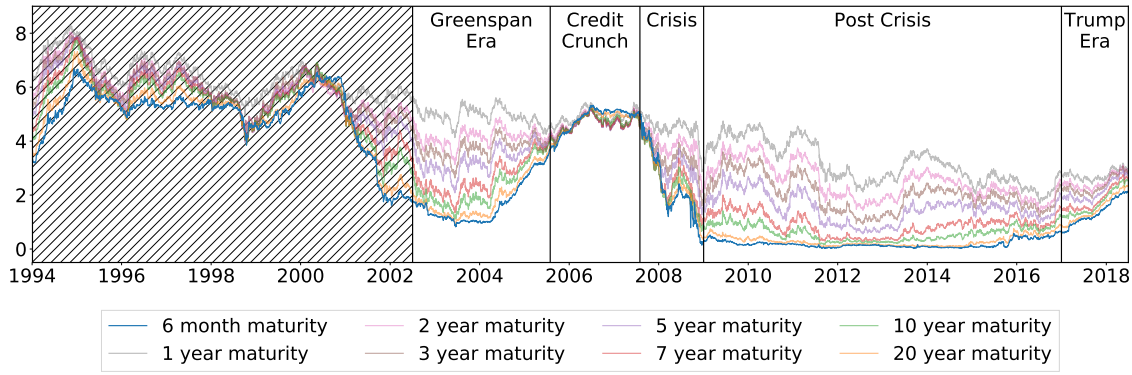
Table 6.2: Summary statistics of the monthly returns / changes

Asset	Mean	Volatility	Skewness	Kurtosis
Exchange rate returns				
AUD	0.0000	0.0368	0.7922	5.9807
CAD	-0.0003	0.0265	0.7218	6.4690
CHF	-0.0011	0.0297	-0.0101	4.8305
EUR	0.0003	0.0292	0.3192	4.1236
GBP	0.0013	0.0252	0.5218	4.9380
JPY	0.0002	0.0281	0.3077	3.5187
NZD	-0.0003	0.0383	0.5727	4.6172
SEK	0.0011	0.0328	0.1555	3.5244
Interest rate changes				
6 month	-0.0039	0.2016	-2.2169	14.6083
1 year	-0.0041	0.2137	-1.2097	8.6316
2 year	-0.0055	0.2462	-0.3439	4.4936
3 year	-0.0062	0.2617	-0.0753	3.9820
5 year	-0.0078	0.2728	0.0288	3.7443
7 year	-0.0086	0.2697	0.1118	3.7856
10 year	-0.0097	0.2594	-0.0196	4.2279
20 year	-0.0115	0.2372	0.0341	4.6910
Commodity index returns				
Copper	0.0065	0.0724	-0.0517	5.8856
Corn	-0.0048	0.0752	0.2992	4.0710
Gold	0.0027	0.0453	0.1885	4.1817
Live Cattle	-0.0006	0.0392	-0.4110	5.1238
Natural Gas	-0.0080	0.1316	0.4827	3.9878
Soybean	0.0047	0.0684	-0.0485	3.5828
Sugar	0.0037	0.0886	0.2235	3.4728
WTI Oil	0.0034	0.0876	-0.0136	3.8353

The monthly return and changes are calculated using the values at the start of each month which the summary statistic aggregates over the time periods specified in Table 6.1. Our study only applies daily data but we use a monthly frequency in this table to avoid minuscule magnitudes.

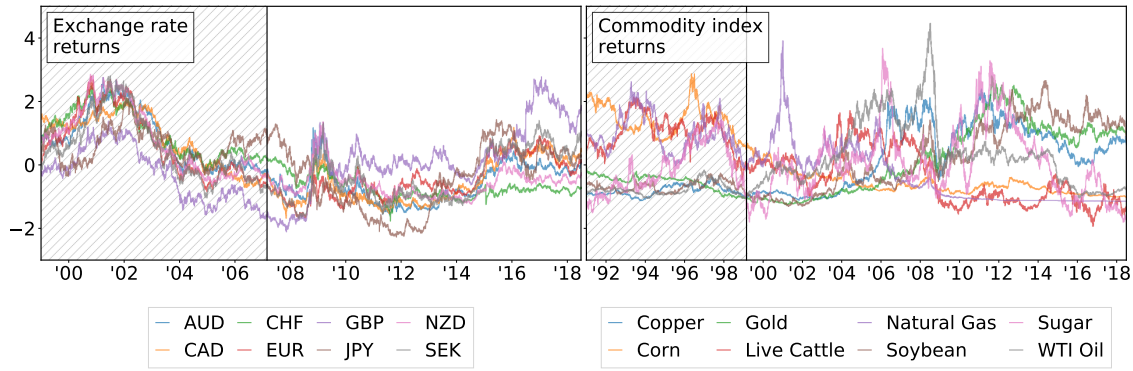
structure is not observed in either the exchange rate or the commodity index data in Figure 6.3.

Figure 6.2: Regimes for US interest rates



The US interest rates are segmented into five regimes which differentiate regarding their properties due to macro-economic influences. The shaded period before July 2002 is only used for initial model calibrations (and models are then re-calibrated daily on a fixed-size moving sample) and so it is excluded from the out-of-sample period.

Figure 6.3: Regimes for exchange rates and commodity indices



We standardize the data such that the mean is zero and the standard deviation is 1 for an easier visualisation. Neither exchange rates or commodity indices showcase any discernible regimes within our sample size. The shaded periods are only used for initial model calibrations (and models are then re-calibrated daily on a fixed-size moving sample) and so they are excluded from the out-of-sample period.

To examine the robustness of our analysis, we segment the data into three parts, ranging from 2006–2010, 2010–2014, and 2014–2018 with breakpoints at the end of June in each case. Because the exchange rate data starts much later than the other data sets, the period from June 2006 to February 2007 is still used for calibration. Therefore, the first sub-period begins in March 2007 in this case. Given

the distinct regimes in the US interest rate data, we additionally examine each of the five sub-periods in Figure 6.2 individually.

Within the exchange rate data, some values emerge from unforecastable and surprising macro-economic events. Figure 6.4 overlaps all eight exchange rate return time series to identify the outliers. These correspond to the following dates and changes relative to USD:

(i) **6 September 2011:** The Swiss Franc is pegged against the Euro to protect the Swiss economy against the European debt crisis, resulting in a 10% increase of the CHF within one day;

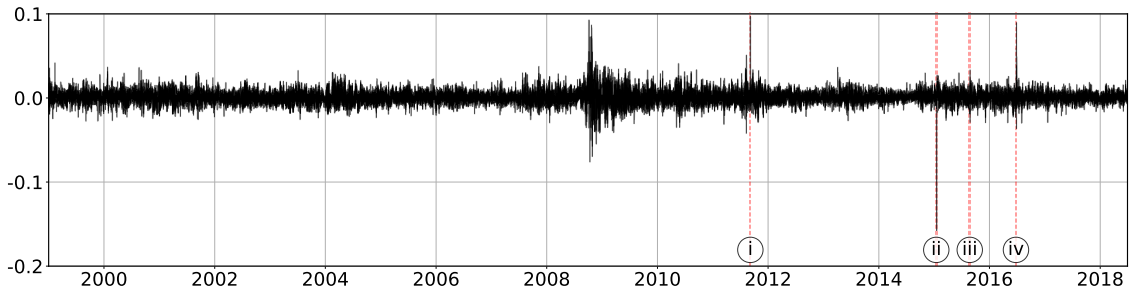
(ii) **15 January 2015:** The Swiss National Bank reverts to a floating exchange rate with the Euro. This devalues the Swiss Franc by 16% in one day;

(iii) **24 August 2015:** Euro exchange rates rise due to the Greek sovereign crisis. Despite only causing a minor decrease of 2%, this is one of the four largest drops for the Euro within a 5-year window.

(iv) **24 June 2016:** The UK votes in a referendum to leave the EU (Brexit). GBP increases by 9%, its largest daily change within our sample period.

The removal of the four outliers is further motivated by the inability of any model in our empirical study to accurately forecast daily returns on these dates.

Figure 6.4: Exchange rate outliers



We overlap all eight time series of the exchange rate returns to identify the outliers within the data set. These are identified with a red line and an associated number for reference. The highly volatile period around 2009 is not classified as an outlier because it corresponds to a systemic change in the market conditions.

We remove all four data points from our accuracy evaluation in Section 6.3 because we do not want any of those unforecastable events, for which superior forecasting performance can only be attributed to chance, to dominate the MCS. However, we keep the highly volatile period around 2009 in our study since a good model should be able to adapt to the changing market conditions.

6.2 Empirical Design

We quantify the relative forecasting performance of Factor Quantile models with the latent specifications in Section 4.3 against several popular benchmark methods through the evaluation methodology outlined in Chapter 5. Five different weightings of CRPS from Table 5.1 measure the univariate accuracy while the multivariate quality is assessed through the energy score and three parametrisations of the variogram score. For simplicity, we do not scale or normalize the scores although proper scoring rules remain proper after such a transformation (Toda, 1963). All scores are calculated based on an ensemble consisting of 100,000 draws from the respective distributions. The use of multiple proper scoring rules is motivated by their different focus and is especially relevant in higher dimensions where the encapsulation into a single score is potentially associated with a large loss of information.

Sequential testing with MCS applies the proper scores as loss functions to rank the models according to their accuracy. Since the distribution of the test statistic $T_{\mathcal{M}}$ in Equation 5.1 is non-standard, it has to be estimated through a bootstrap procedure. To this end we employ a block–bootstrap with 5,000 bootstrap replications and a block-length that is determined by the maximum number of significant parameters during the fitting of an autoregressive model on the relative performance variable.

Our empirical study includes a parsimonious set of benchmarks and Factor Quantile parametrisations with independent marginals as well as models built on empirical correlations. This allows us to test the sensitivity of different multivariate scoring rules to assumptions about correlation.

The first type of benchmark models are CCC- and DCC-GARCH with E-GARCH volatilities and Student-t innovations. These models capture salient properties of financial time series i.e. volatility clustering, skew and heavy tails and asymmetric response to shocks. At the same time, they are easier to calibrate than GARCH models with mixture normal distributions or other, more complicated features. Both multivariate GARCH models are calibrated using maximum likelihood estimators adapted from the 2-stage implementation in the Oxford MFE Toolbox by Sheppard

(2013) to utilize E-GARCH with Student-t distributed innovations. That is, we have replaced the univariate Gaussian GARCH(1,1) for CCC- and DCC-GARCH with Student-t E-GARCH(1,1). This choice is motivated by Hansen and Lunde (2005) who provide an extensive comparison of 330 univariate GARCH specifications through the Hansen (2005) superior predictive ability data-snooping check, concluding that it is hard to beat an asymmetric GARCH(1,1) model with Student-t innovations.

Our second type of benchmark model is the empirical distribution function (EDF) with either independent marginals or a Gaussian copula using a historical correlation matrix which is estimated on the same data used for calibration. This copula model can be easily applied in high-dimensional systems and performs well in previous forecasting exercises (Patton, 2012, 2013). There are, of course, numerous alternative parametric choices for both marginals and copula, as described by Patton (2013). But we have over 96,000 distribution forecasts to generate in total for each model, and this number of high dimensional calibrations for more complicated parametric copulas is not feasible. By the same token, we only consider the Gaussian copula because robust estimation of parameters even for 8-dimensional parametric copulas is too great a computational challenge for an exercise of this scale. Using EDF marginals based on the same data as the Factor Quantile marginals additionally allows us to test the effectiveness of PCA factor models, in the context of quantile regressions, for reducing the noisy variation which could deteriorate forecasting accuracy of models with EDF marginals.

We do not include the random walk model although this is a common benchmark in exchange-rate forecasting. This is because it does not yield a distribution forecast and as such is no alternative to Factor Quantile models.

Regarding the Factor Quantile specifications, we apply the latent versions based on the last principal components (FQ-AL) from Section 4.3.1 and asymptotic bagging (FQ-AB) from Section 4.3.2 with either independent marginals or the same Gaussian copula as the EDF. Both specifications of our Factor Quantile model use the quantile partition

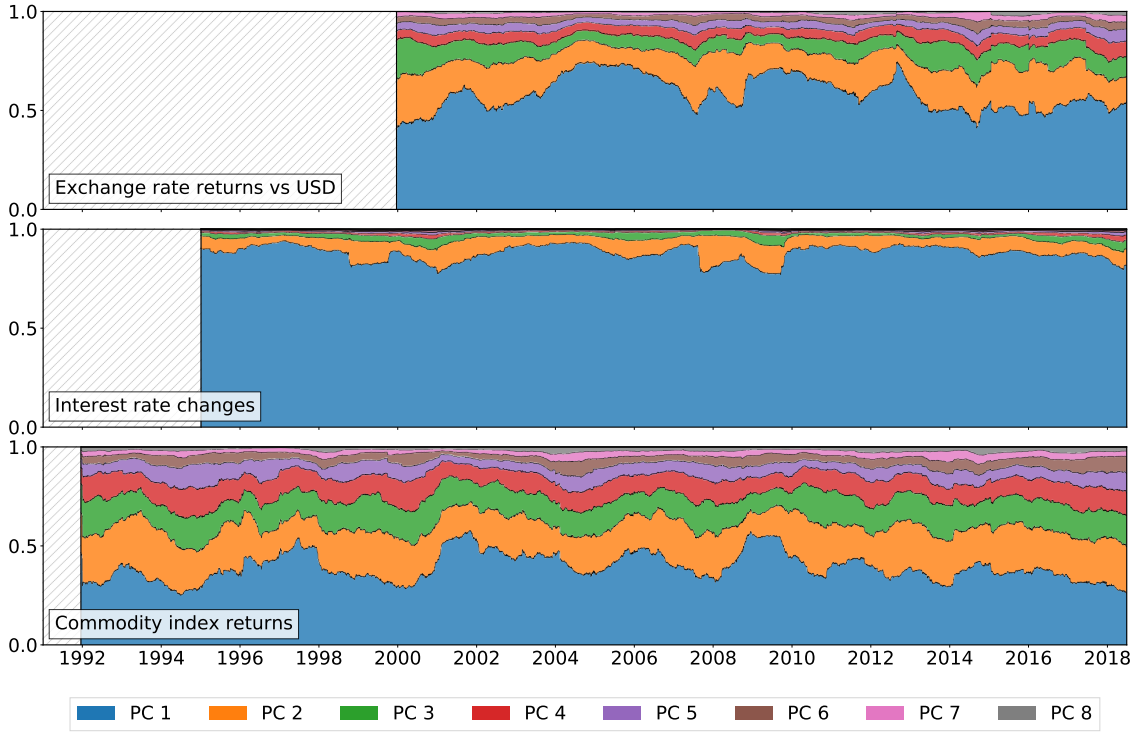
$$\mathbb{Q}_9 = \{0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\}$$

for the regressions and employ the shape-preserving method for interpolating distribution functions. There is no guarantee that the conditional quantiles of FQ-AB are monotonic and we refer to estimation methods by Koenker (2005) or Chernozhukov et al. (2010) to circumvent this issue. Nevertheless, during the entire course of our empirical study, the estimated conditional quantiles exhibited no crossing behaviour on any data set with any of the calibration choices, indicating that our factor models are well-conditioned. Figure 6.5 illustrates how the number of principal factors m is selected, by depicting the cumulative variance explained by the rolling principal components over the available data period for each asset class. The number of components needed to capture most of the variance exhibits distinct patterns. In FQ-AB, we select $m = 4$ components as common factors for the exchange rates, $m = 2$ for the interest rates and $m = 6$ for the commodity indices. On average, over the entire period shown, together the four components explain 90% of the variation in the exchange rate data, the two components explain 95% of the variation in the interest rates, and the six components explain 95% of the variation in the commodity returns. Following the same reasoning, FQ-AL uses $m = 4$ components as common factors for the exchange rates, $m = 6$ for the interest rates and $m = 2$ for the commodity indices.

We avoid data snooping by using a broad range of data sets with assets motivated through economic factors rather than the predictive prowess of our models. All parameters of the Factor Quantile models are chosen based on criteria that are available ex-ante. Additionally, we quantify the performance based on very long time series, further limiting the probability that any superior performance can be attributed to chance.

It is well known that multivariate GARCH models can have ill-conditioned likelihood functions which are hard to optimize unless the calibration sample has sufficient size, so we have selected 2,000 daily returns for the calibration of each time

Figure 6.5: Cumulative variance explained by the principal components



The variance explained is based on rolling principal components for the three data sets. We use 250 observations for the calculation of the covariance matrices.

Table 6.3: Summary of models used in the empirical study

Model	Marginals	Dependency	Calibration
FQ-AL$^{C}_{250}$	Alpha FQ w/ last PC	Gaussian copula	250
FQ-AL$^{C}_{2000}$	Alpha FQ w/ last PC	Gaussian copula	2,000
FQ-AL$^{I}_{250}$	Alpha FQ w/ last PC	Independent	250
FQ-AL$^{I}_{2000}$	Alpha FQ w/ last PC	Independent	2,000
FQ-AB$^{C}_{250}$	Asym. Bagging FQ	Gaussian copula	250
FQ-AB$^{C}_{2000}$	Asym. Bagging FQ	Gaussian copula	2,000
FQ-AB$^{I}_{250}$	Asym. Bagging FQ	Independent	250
FQ-AB$^{I}_{2000}$	Asym. Bagging FQ	Independent	2,000
EDF$^{C}_{250}$	EDF	Gaussian copula	250
EDF$^{C}_{2000}$	EDF	Gaussian copula	2,000
EDF$^{I}_{250}$	EDF	Independent	250
EDF$^{I}_{2000}$	EDF	Independent	2,000
CCC-GARCH	Student-t E-GARCH(1,1)	Conditional correlation	2,000
DCC-GARCH	Student-t E-GARCH(1,1)	Dyn. cond. correlation	2,000

We compare multivariate GARCH models and traditional Gaussian copulas with EDF marginals against our two latent Factor Quantile models. To capture the long-term variance, our GARCH models use a long calibration window. For the copula models, we use a correlation matrix derived from historical estimation with the same calibration length as the marginals or an identity matrix.

series. For consistency with the GARCH models, we have also taken 2,000 data points for the quantile regressions. However, we have found that Factor Quantile works well with fewer data points than GARCH models; indeed quantile regression yields robust estimates with principal component factors even with a sample size of 250. To avoid taking too many Factor Quantile models forward for comparison with the GARCH and EDF benchmarks, we have therefore selected to present results for sample sizes of 2,000 and 250.[2] For the EDF marginals with Gaussian copula, we also choose calibration sample sizes of both 250 and 2,000. The marginals use the same calibration sample as the Gaussian copula.[3] Table 6.3 summarises the set of benchmark models and the Factor Quantile parameterisations that we apply in the remainder of this study.

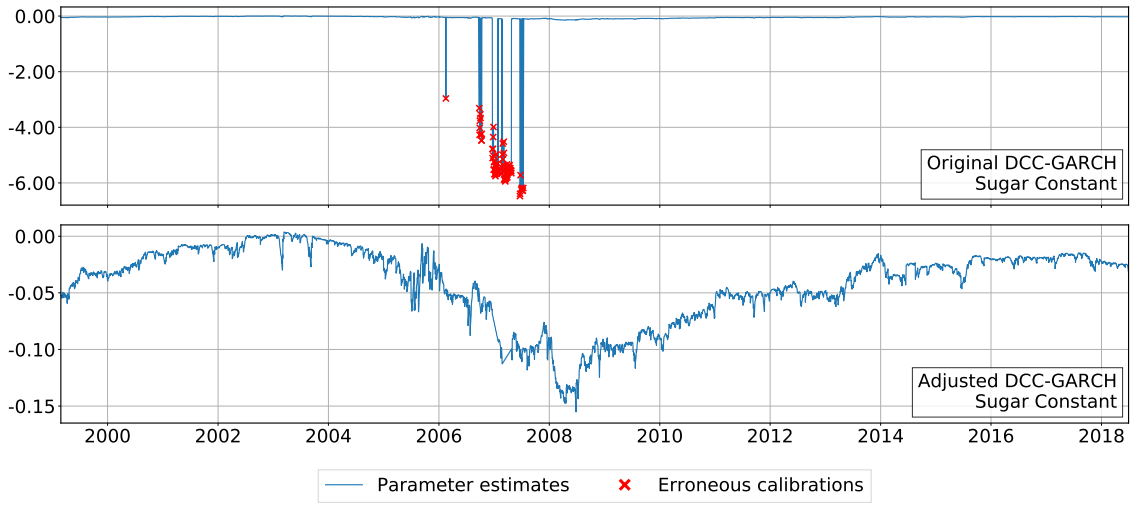
All models are re-calibrated daily with only data available up to that time to avoid forward-looking bias. The estimated parameters are subsequently used to generate one-day-ahead distribution forecasts. Then the fixed-size calibration sample is rolled forward one day and the forecasts are repeated. In total we estimate each multivariate model around 12,000 times and with 14 different models and eight dimensions this yields more than 1.3 million distribution forecasts for further analysis. We compare the resulting scores both for the entire out-of-sample period and for specific sub-periods to evaluate the robustness of the forecasting performance over time.

Despite the long calibration period, both multivariate GARCH models exhibit issues with parameter calibration during our empirical study because the likelihood functions become challenging to optimize with eight dimensions. For instance in the commodities data, the constant and GARCH parameters do not always converge to sensible values for the live cattle marginal, and many parameters converge

[2]Performances for alternative calibrations are excluded but are available upon request. The flexibility of Factor Quantile models in terms of calibration choice is one of its advantages, making the methodology amenable to a wide variety of time series data.

[3]As discussed in Section 5.1.2, the MCS elimination rule excludes models that are significantly inferior to other ones in the initial set $\mathcal{M}^0$. Therefore, it is likely that model variations with alternative calibration sizes remove each other from the superior set due to their similarity. This further motivates our choice to only consider two calibration lengths that cover short and long estimation periods respectively.

Figure 6.6: Convergence issues with GARCH models (sugar)



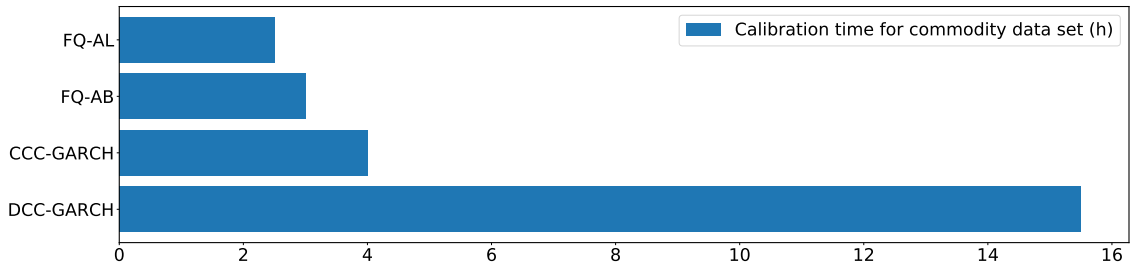
The parameter illustrated is the sugar constant parameter estimated for DCC-GARCH. The upper figure shows the parameter obtained using the adapted Oxford MFE toolbox and the lower figure shows the parameter after replacing erroneous calibrations with the most recent unproblematic value. Parameters that differ by a very large amount from previous estimations are classified as mis-calibrations.

to unrealistic values for sugar. These mis-calibrated parameters require manual attention, which prevents full automation of multivariate GARCH models. Our model accuracy tests therefore exchange erroneous parameters with the most recent unproblematic values, as illustrated by Figure 6.6. The mis-calibration might also be avoided by replacing the maximum likelihood estimation with a more advanced one based on Markov Chain Monte Carlo (MCMC) but this would further increase the computation time of multivariate GARCH models. Karlsson (2013) urges for a careful assessment of the convergence of the posterior distribution which is especially relevant in multivariate settings with high degrees of freedom.[4]

It is worth noting that Factor Quantile models are much faster to calibrate than multivariate GARCH, even without dealing with any of the latter's convergence issues. For instance, daily re-calibration over a rolling window on the data set with eight commodity sub-indices yields the computation times illustrated in Figure 6.7.

[4]We refer to Ardia (2008) for a discussion on MCMC for univariate GARCH calibrations. Virbickaite et al. (2015) surveys various Bayesian implementations in an univariate and multivariate setting. Of particular relevance is the Bayesian approach by Virbickaitė et al. (2016) that can be applied to asymmetric DCC-GARCH models. Asai (2006) compares the computational efficiency of several MCMC methods, including the Metropolis-Hastings algorithm and the greedy Gibbs sampler.

Figure 6.7: Comparison of calibration time (commodity index returns)



The calibration time is measured on an Intel i5-6500 with 3.20 GHz. Over 40,000 daily forecasts, i.e. 5,046 for each of the eight commodity index returns, are generated with each model. All timings are for models with calibration on 2,000 observations.

This makes Factor Quantile models at least 30% faster than CCC-GARCH and more than five times faster than DCC-GARCH. Note that the current implementation of Factor Quantile models is based on Python while the multivariate GARCH models use optimized MATLAB functions. As the efficiency of MATLAB is generally higher than that of Python scripts, we expect that the difference in speed would become even more pronounced when comparing the multivariate GARCH models to an optimized Factor Quantile algorithm.

6.3 Forecasting Accuracy Results

We segment the accuracy results into the univariate and the multivariate forecasting performance and examine each over the entire sample period as well as over sub-periods. As Diebold (2015) argues, the relative performance of competing models should be examined using all available data. The evaluation on the sub-periods is primarily an analysis on the robustness of the scoring rules rankings that augments the full-sample accuracy discussion. Both Gneiting et al. (2008) and Scheuerer and Hamill (2015) emphasize the importance of testing the marginal distributions. Applying multivariate tests alone is not sufficient because we require a model that forecasts accurate marginals as well as one that correctly captures the dependence between them.

The MCS approach analyses the performance of both Factor Quantile models separately against the benchmark models since we aim at quantifying the accuracy of each individual Factor Quantile specification. In the sequential hypothesis test, models get removed from the superior set of models if they are inferior to another model given some confidence level. Therefore, a MCS analysis with both FQ-AL and FQ-AB in the initial set $\mathcal{M}^0$ may exclude some of our models that perform well individually but are overshadowed by the better Factor Quantile specification. We only report the results for $\alpha = 0.25$ but the findings for all other confidence levels can be extracted from our MCS tables – see Section 5.1.2 for a discussion.

Section 6.3.1 presents the results of the CRPS. Since the choice of copula does not impact the marginals, some of the models are identical with respect to their univariate accuracy and we end up with 8 competing models in the comparison. For ease of notation, we drop the superscripts C and I of Table 6.3 during this discussion. In addition to MCS, we apply the CRPS test statistic described in Section 5.2.1 to obtain a more detailed comparison and to verify the MCS results.

Multivariate accuracy is discussed in Section 6.3.2 where we apply the energy score and the variogram score with $p = 0.5, 1, 2$. This is, to the best of our knowledge,

the first extensive application of these multivariate scoring rules in finance and economics.

We only include the most relevant tables and figures in Sections 6.3.1 and 6.3.2. The more detailed results are numerous, and are available in Appendix A or as supplementary materials electronically.

6.3.1 Univariate Forecasting Accuracy

We start the univariate evaluation with the MCS results based on the overall CRPS. These rankings utilize the entire out-of-sample size in Table 6.3 which includes at least 3,000 out-of-sample observations in each data set. Tables 6.4 and 6.5 list the p-values of the MCS for uniformly weighted CRPS. All other tables corresponding to different weights can be found in the appendix. Depending on the individual assets and the data, the performance of each model varies strongly. This further emphasizes the importance of the data and ex-ante asset selection in Section 6.1 since otherwise assets could be chosen to favour certain models.

Generally, FQ-AL performs very well, being either the model with the most or second most inclusions in the superior set of models. This is particularly promising, since the best model changes for each data set, making FQ-AL the most accurate model overall. Generally, we observe that models based on 250 observations almost always outperform their counterparts with 2,000 observations. This may be explained by a changing data generating process over time to which models with long calibration windows cannot adapt quickly enough.[5] We hence focus our discussion on the Factor Quantile models with a 250 calibration window.

Exchange rate returns The uniformly weighted CRPS identifies CCC-GARCH as the most accurate model and includes it in 63% of the superior sets. Our Factor Quantile specification FQ-AL_{250} follows as the second best model with an inclusion rate of 38%. This gap closes when all the five weighted CRPS are considered, resulting in 58% and 45% for CCC-GARCH and FQ-AL_{250} respectively.

Interest rate changes FQ-AL_{250} dominates this data set and remains in 75% of the superior sets. This is around 3 times higher than that of EDF_{250}, the next best model. These results are robust and remain valid for the uniformly

[5]If this is indeed the case, it mostly affects the Factor Quantile and EDF models since both CCC-GARCH and DCC-GARCH apply a conditional covariance structure which mitigates the issue. Further, both GARCH models are restricted to long calibration periods for the estimation of the long-term variance and the stability of calibrated parameters.

weighted CRPS as well as for all five CRPS weights. The accuracy of Factor Quantile models for interest rate changes is expected since the data is highly correlated which benefits our application of PCA.

Commodity index returns Multiple models perform well in the commodity data set. The uniformly weighted CRPS keeps FQ-AL_{250} and DCC-GARCH in 38% of the superior sets. CCC-GARCH follows closely with 25%. A clearer ranking forms when the weighted CRPS is considered. In this case, DCC-GARCH becomes the best model with a 45% inclusion rate compared to 33% for FQ-AL_{250}.

The alternative Factor Quantile specification FQ-AB performs similarly to FQ-AL but is more accurate, especially for commodity index returns. As mentioned in Chapter 4, the Factor Quantile bagging algorithm considers the entire distribution of the conditional quantiles rather than focusing only on the their expectation. This may explain the better performance. Relative to FQ-AL, the accuracy of FQ-AB based on uniformly weighted CRPS changes as follows:

(i) FQ-AB_{250} replaces CCC-GARCH as the best model for EUR, JPY, and oil as well as replaces EDF_{250} as the best model for gold. Additionally, the Factor Quantile model is also represented in the MCS for live cattle.

(ii) However, the performance for copper weakens and the MCS contains additionally to FQ-AB_{250} also DCC-GARCH.

(iii) Furthermore, FQ-AB is less accurate for CHF where it is substituted by DCC-GARCH;

Based on the uniformly weighted CRPS, the inclusion rates for FQ-AL_{250} and FQ-AB_{250} stay the same in exchange rate returns and interest rates changes but increases to 63% in commodity index returns. The change is even more pronounced in the weighted CRPS results, where FQ-AB becomes the best model for all three data sets. GARCH models retain their relatively good performances and are the

Table 6.4: MCS p-values for FQ-AL: Uniformly weighted CRPS

Asset	GARCH		EDF		FQ-AL	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.00	1.00**	0.00	0.06	0.00	0.00
CAD	1.00**	0.00	0.00	0.08	0.00	0.00
CHF	0.00	0.15	0.00	0.00	1.00**	0.00
EUR	0.29**	0.00	0.00	0.00	0.21*	1.00**
GBP	0.00	1.00**	0.00	0.00	0.00	0.00
JPY	1.00**	0.00	0.00	0.00	0.30**	0.00
NZD	1.00**	0.00	0.00	0.00	0.00	0.00
SEK	0.50**	0.00	0.00	0.00	1.00**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.00	0.00	0.01
1 year	0.00	0.00	1.00**	0.00	0.00	0.00
2 year	0.00	0.00	0.95**	0.00	1.00**	0.00
3 year	0.00	0.00	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.00	1.00**	0.00
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.16*	0.16*	0.00	0.00	1.00**	0.00
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.01	0.00	1.00**	0.00	0.00	0.00
Live Cattle	0.00	1.00**	0.00	0.00	0.00	0.00
Natural Gas	0.00	1.00**	0.00	0.00	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	0.86**	0.00	0.00	0.00	1.00**	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue. Corresponding tables for other weights are in the appendix.

Table 6.5: MCS p-values for FQ-AB: Uniformly weighted CRPS

Asset	GARCH		EDF		FQ-AB	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.02	1.00**	0.02	0.02	0.01	0.23*
CAD	1.00**	0.00	0.12*	0.00	0.00	0.12*
CHF	0.01	1.00**	0.00	0.00	0.00	0.01
EUR	0.00	0.00	0.00	0.00	1.00**	0.00
GBP	0.04	1.00**	0.00	0.00	0.00	0.00
JPY	0.00	0.00	0.00	0.00	1.00**	0.06
NZD	1.00**	0.03	0.00	0.00	0.00	0.00
SEK	1.00**	0.00	0.00	0.00	0.48**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.00	0.00	0.01
1 year	0.07	0.00	1.00**	0.00	0.00	0.00
2 year	0.09	0.00	0.00	0.00	1.00**	0.00
3 year	0.00	0.00	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.00	1.00**	0.00
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.00	1.00**	0.00	0.00	0.82**	0.16*
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.24*	0.01	0.00	0.00	1.00**	0.01
Live Cattle	0.00	0.33**	0.00	0.00	1.00**	0.00
Natural Gas	0.00	1.00**	0.00	0.00	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	0.00	0.10*	0.00	0.00	1.00**	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue. Corresponding tables for other weights are in the appendix.

second best models in exchange rate returns and commodity index returns where the respective GARCH specification lags 10% and 7% behind FQ-AB_{250} respectively. While the percentage inclusion in interest rates does not increase between FQ-AL_{250} and FQ-AB_{250}, the latter model manages to reduce the number of EDF models in the superior sets by being significantly more accurate.

This sets up FQ-AB_{250} as a very promising alternative to our benchmark models but also shows that it behaves differently to FQ-AL_{250}. The change in the underlying factor model, coupled with the bagging approach that accounts for the entire distribution of the conditional quantiles leads to deviations in the forecasted distribution, similar to the 6-month interest rate example in Section 4.3.2.

Notably, both Factor Quantile versions never manage to be included in the superior set for the interest rates with 6-month and 1-year maturity. This is likely because these interest rates at the short end are policy instruments and mostly influenced by money market operations. In contrast, interest rates with longer maturities depend largely on swaps. Therefore, the underlying dynamics behind the interest rates with 6-month and 1-year maturity differ from the others and may not be well described by the principal component representation.

Furthermore, the performance of the empirical distribution function is much worse than that of both Factor Quantile models. In fact, it only performs well in interest rates with 1- or 2-year maturities and gold. In the comparison with FQ-AL, EDF models are additionally included for the right-tail weighted CRPS for live cattle and the centre weighted CRPS for 6-month maturity interest rates. This indicates that the principal component representation succeeds at reducing the noise of the observed historical data and produces significantly more accurate forecasts.

The weights of the CRPS only play a secondary role in the evaluation. Slight deviations to the uniformly weighted CRPS case are present, but only in a relatively low amount of cases – 23% in exchange rate returns, 6% in interest rate changes and 17% in commodity index returns for the comparison with FQ-AL and 17% in exchange rate returns, 2% in interest rate changes and 19% in commodity index

returns for the comparison with FQ-AB. The particularly stable results in the term structure data can be explained by the vast out-performance of the Factor Quantile models in this data set. Changes are mostly limited to one weight only, with two notable exceptions:

(i) The DCC-GARCH model for CHF and the FQ-AL_{250} model for EUR are included for the other four weighted CRPS cases but not for the uniformly weighted one. This is surprising but not impossible since the weights transform the CRPS to focus on different parts of the distribution but their outcomes cannot be combined to yield the uniformly weighted CRPS.

(ii) This also happens to a lesser degree for DCC-GARCH in copper. It is not represented in the superior set of the uniformly and right-tail weighted CRPS but in all three sets corresponding to the remaining weighted CRPS.

All other changes are limited to one or two weights only. For the FQ-AL comparison, 68% of the cases where there is a deviation from the uniformly weighted CRPS are constrained to a single alteration and 18% to two alterations. The FQ-AB comparison shows a similar pattern with 61% and 33% respectively.

Interestingly, the MCS tables show that DCC-GARCH is not always better than CCC-GARCH. The benefit of DCC-GARCH over CCC-GARCH is its time-varying correlation but this relates to the dependency structure and does not translate to a superior univariate performance. Only the commodity data set shows consistent improvements compared to CCC-GARCH.

The superior set of models may include more than one model in the case that the forecasting accuracy of the remaining models cannot be distinguished with the equivalence test given a pre-specified confidence level. However, in most of our tests MCS identifies a single model as the superior one. This suggests that our out-of-sample period is informative enough to select a best model unequivocally.

Table 6.6 shows the percentage of times that FQ-AL_{250} and FQ-AB_{250} beat each of the four benchmark models significantly based on the asymptotically normal CRPS

test statistic described in Section 5.2.1 using the entire out-of-sample period. In contrast to the MCS tables, this classic hypothesis test can only compare two models directly. Hence, it yields less informative results but does not rely on any bootstrap estimation for the distribution of the test statistic contrary to MCS. We use this second test to validate the MCS results and to obtain a more detailed view on the performance of FQ-AL_{250} and FQ-AB_{250}.

Table 6.6: Summary of CRPS hypothesis tests

Model	FQ-AL_{250}			FQ-AB_{250}		
	50%	95%	99%	50%	95%	99%
Exchange rate returns						
CCC-GARCH	58%	45%	40%	70%	58%	58%
DCC-GARCH	50%	48%	48%	55%	50%	48%
EDF_{250}	90%	83%	83%	88%	88%	88%
EDF_{2000}	100%	88%	88%	93%	88%	88%
Interest rate changes						
CCC-GARCH	88%	85%	85%	88%	88%	85%
DCC-GARCH	88%	88%	88%	88%	88%	88%
EDF_{250}	68%	63%	63%	75%	70%	70%
EDF_{2000}	100%	100%	98%	98%	98%	98%
Commodity index returns						
CCC-GARCH	53%	43%	43%	78%	68%	65%
DCC-GARCH	45%	40%	40%	60%	53%	53%
EDF_{250}	75%	73%	73%	100%	98%	90%
EDF_{2000}	75%	75%	75%	100%	100%	100%

This table shows the percentage of times that FQ-AL_{250} and FQ-AB_{250} beat the alternative model in a hypothesis test based on the CRPS test statistic at the p-value listed in the column heading. We consider all assets and all five CRPS weights. The 50% threshold is included to shows how often a benchmark model may be favoured over the Factor Quantile Models. All tables with the individual p-values for all weights and assets can be found in the supplementary materials.

The results largely agree with those of MCS but include some details on the performance of our Factor Quantile specification in case it is not in the superior set, or in case there are multiple models in the superior set. Because the combinations between CRPS weights, asset and model pairs lead to a large amount of hypothesis tests, the probability of some false positives is high.[6] The CRPS results should therefore be viewed as only supplementary to the MCS results. In summary:

[6]Of course, this is partly mitigated by the fact that we average over multiple hypothesis test applications. Each cell in Table 6.6 represents the result of 40 tests.

(i) Even if our Factor Quantile models are not in the superior set, they usually are the second best model that manages to beat all benchmark models but one. There are a few exceptions to this, where we do not have a significantly higher accuracy than several other models. For FQ-AL_{250}, this is the case for CAD, GBP, 6-month interest rate, copper, gold, and live cattle. Similarly, FQ-AL_{250} is either the best or second best model apart from CAD, CHF, 6-month interest rate and copper.

(ii) We also find one instance where the CRPS test statistic slightly deviates from the MCS. FQ-AL_{250} does not manage to significantly outperform either GARCH model for copper but is represented in the superior sets of all five weights in comparison to CCC-GARCH that remains in one, and DCC-GARCH that remains in four sets. This may result either from the estimation of the test statistic distribution through bootstrap in MCS or errors in the CRPS hypothesis tests.

We now consider the MCS results over time to assess the robustness of our results. Table 6.7 summarizes the inclusion rates for each model in three sub-periods. Tables with the sub-sample p-values are available in the supplementary materials.

The three sub-periods indicate that the accuracy of the marginal forecasts may change drastically over time. However, generally the model that performs best for the entire sample period does not change within the sub-periods. There are two exceptions to this:

(i) In exchange rate returns, Factor Quantile models are particularly good for the two most recent sub-periods and manage to beat CCC-GARCH in terms of accuracy. This may be the effect of the financial crisis that remains in the calibration window for the GARCH models through the remainder of the evaluation. If this is the case, it again represents a structural issue with the calibration requirements of GARCH models with no adequate direct solutions since removing the financial crisis yields a fragmented time series that may be unfit for statistical analysis.

Table 6.7: Comparison of univariate performance over time

Sample	GARCH		EDF		FQ	
	CCC	DCC	250	2000	250	2000
FQ-AL comparison: Exchange rate returns						
All	58%	35%	0%	0%	45%	8%
2007 to 2010	55%	35%	3%	10%	25%	8%
2010 to 2014	35%	30%	3%	0%	60%	0%
2014 to 2018	20%	30%	5%	5%	65%	20%
FQ-AL comparison: Interest rate changes						
All	13%	0%	28%	3%	75%	0%
2006 to 2010	13%	0%	8%	3%	83%	5%
2010 to 2014	13%	10%	43%	0%	58%	0%
2014 to 2018	15%	13%	30%	13%	58%	8%
FQ-AL comparison: Commodity index returns						
All	25%	45%	13%	0%	33%	0%
2006 to 2010	45%	60%	15%	18%	28%	5%
2010 to 2014	30%	60%	18%	23%	48%	13%
2014 to 2018	38%	63%	10%	3%	23%	8%
FQ-AB comparison: Exchange rate returns						
All	43%	40%	0%	0%	53%	0%
2007 to 2010	53%	43%	5%	5%	38%	23%
2010 to 2014	43%	50%	15%	13%	63%	0%
2014 to 2018	35%	23%	5%	8%	73%	8%
FQ-AB comparison: Interest rate changes						
All	13%	0%	18%	0%	75%	0%
2006 to 2010	13%	0%	25%	18%	83%	0%
2010 to 2014	15%	3%	50%	0%	55%	0%
2014 to 2018	13%	13%	25%	0%	85%	0%
FQ-AB comparison: Commodity index returns						
All	23%	48%	8%	0%	55%	0%
2006 to 2010	48%	63%	15%	13%	50%	38%
2010 to 2014	30%	50%	13%	10%	65%	0%
2014 to 2018	30%	50%	13%	0%	53%	8%

This table shows the proportion of cases that each model is included in the MCS with $\alpha = 0.25$. The best model is highlighted in blue. Each model can be included up to 40 times since we test 8 variables with 5 weightings in each of the three asset classes. We divide the total available backtesting period into sub-periods with breakpoints at the end of June for every year. An exception is the first sample of exchange rates which starts in March 2007. Tables with the underlying MCS p-values are in the supplementary materials.

(ii) In the comparison with FQ-AB, CCC-GARCH and DCC-GARCH have the most inclusions in superior sets during 2007 – 2010 for exchange rate returns and during 2006–2010 for commodity index returns respectively. In all other sub-periods FQ-AB_{250} remains in more superior sets.

It is important to point out that a good relative accuracy during many or even all sub-periods does not guarantee a high percentage of inclusion in the entire out-of-sample period. This is because the MCS has more data during the aggregated periods and is therefore able to exclude models from the superior sets with higher confidence. Additionally, the out-of-sample period 1999 to 2006 is not represented in any sub-period of the commodity index returns.

Notably, the performance of the EDF models rises drastically in the sub-periods for interest rate changes, coming close to the accuracy of Factor Quantile models. This is especially the case for the sub-periods 2010–2014 where EDF_{250} reaches 43% and 50% in the FQ-AL and FQ-AB comparisons respectively. A closer look at the regimes in Figure 6.2 indicates that the interest rates in this period were mostly flat which benefits the historical estimation. A second analysis that segments the sub-periods according to the regimes of the interest rates provides a more detailed view in Table 6.8.

The EDF models are particularly strong in the post crisis and Trump era and even manage to beat FQ-AL_{250} in the latter sub-period. An explanation may be that during these steady times, the time series of US interest rates did not contain as much noise as in prior, more volatile years, eroding the advantages of Factor Quantile models. In contrast, the Factor Quantile models show especially good performance during the credit crunch and crisis. This may be because the interest rates became more correlated which benefits the principal component representation.

Overall, the sub-periods show that the forecasting accuracy varies strongly over time. This is especially true for exchange rate returns, which exhibit the largest ranges for inclusion rates besides the fluctuation for EDF models in interest rates mentioned above: CCC-GARCH ranges from 20% to 55%, DCC from 23% to 50%,

Table 6.8: Performance over different regimes of US interest rates

Sample	GARCH		EDF		FQ	
	CCC	DCC	250	2000	250	2000
Factor Quantile with last principal components						
Greenspan era	8%	10%	5%	13%	85%	0%
Credit crunch	18%	0%	20%	5%	78%	3%
Crisis	20%	3%	10%	15%	83%	3%
Post crisis	20%	0%	33%	0%	58%	0%
Trump era	15%	15%	58%	25%	55%	0%
Factor Quantile with asymptotic bagging						
Greenspan era	10%	10%	33%	13%	88%	0%
Credit crunch	13%	3%	5%	0%	90%	0%
Crisis	23%	3%	0%	0%	93%	0%
Post crisis	25%	0%	40%	0%	53%	0%
Trump era	13%	15%	55%	20%	95%	0%

This table shows the proportion of cases that each model is included in the final MCS with $\alpha = 0.25$. The best model is highlighted in blue. Each model can be included up to 40 times since we test 8 variables with 5 weightings in each of the three asset classes. We divide the total available backtesting period into sub-periods depicted in Figure 6.2. Tables with the underlying MCS p-values are in the supplementary materials.

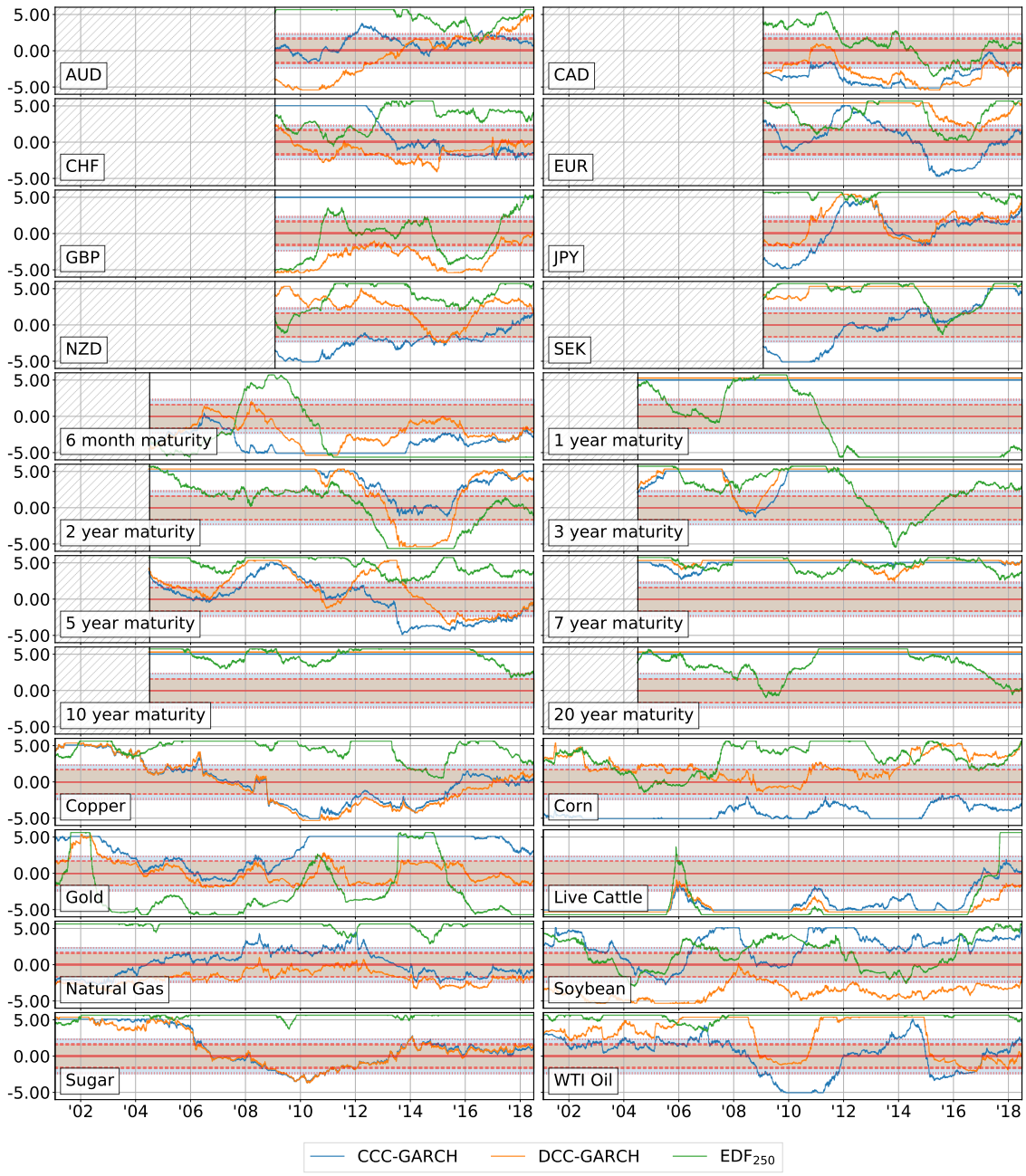
FQ-AL_{250} from 25% to 65% and FQ-AB_{250} from 38% to 73%. Figures 6.8 and 6.9 show the changing performance in even greater detail by plotting the CRPS test statistic based on a rolling window with 500 observations. Each line shows how the respective benchmark model compares against FQ-AL_{250} or FQ-AB_{250} with positive values indicating favourable performance of our Factor Quantile models. The red and blue areas cover (-1.65, 1.65) and (-2.33, 2.33) which means that values beyond them are significant at 5% and 1%. For a clear graphical representation, we limit the comparison to the three best benchmark models and cap the test statistic values at 5, 5.3, and 5.6 for CCC-GARCH, DCC-GARCH and EDF_{250}. This is justified since any value with magnitude above 2.33 is already highly significant:

(i) As indicated in Table 6.8, Factor Quantile models behave much better than EDF models during the financial crisis. Even for the interest rates with 6-month and 1-year maturity where both FQ-AL_{250} and FQ-AB_{250} never are represented in the superior sets, their relative performances are significantly better than

EDF_{250} around 2008. Again, this may be due to the higher correlation between the interest rates that facilitates a better principal component representation.

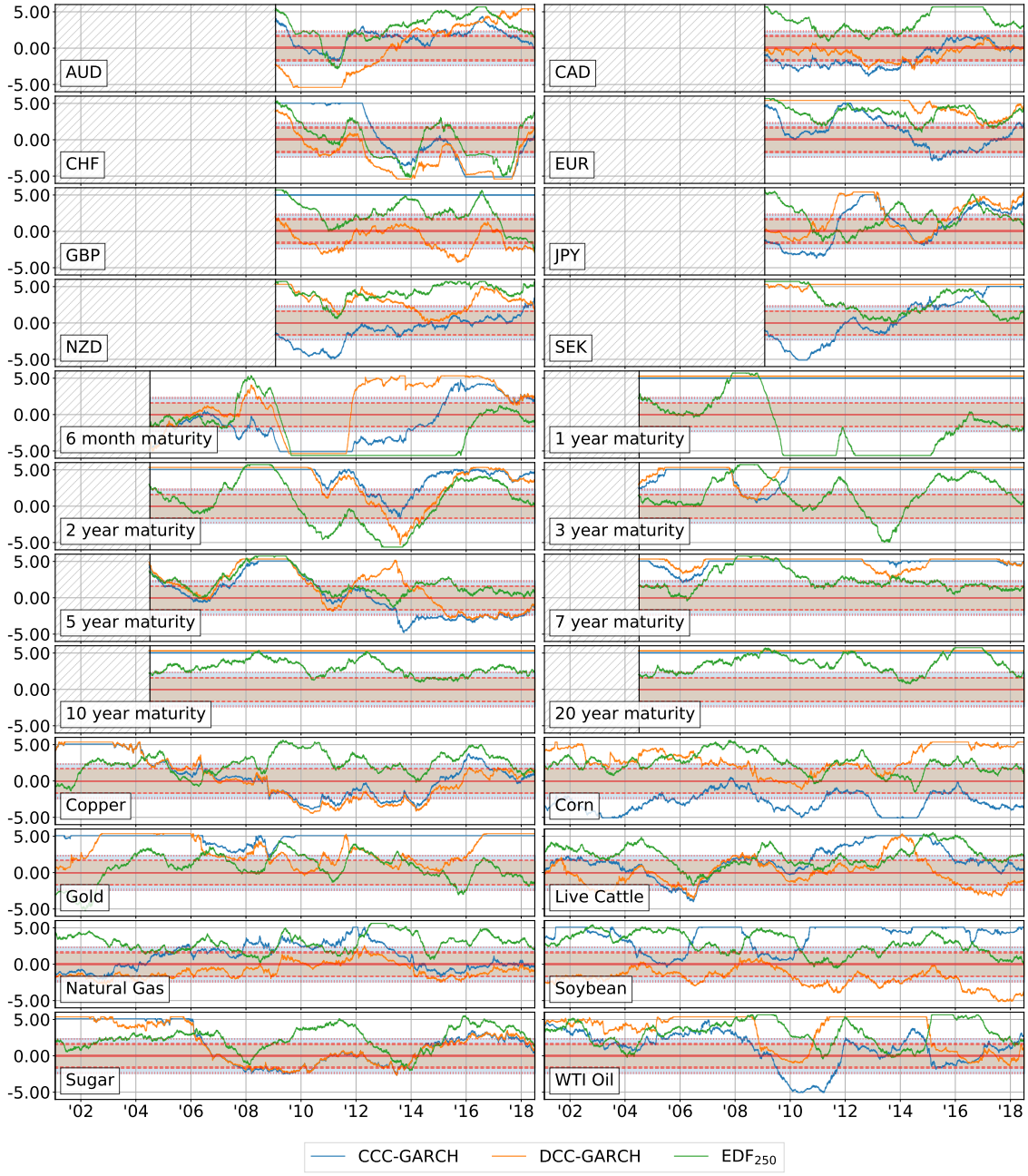
(ii) Even with 500 observations, the CRPS test statistic varies strongly over time, with most models being significantly worse or better than the Factor Quantile models at some point. This, again emphasizes the need for long out-of-sample testing to get accurate and robust results.

Figure 6.8: FQ-AL$_{250}$: Uniformly weighted CRPS test statistic



We plot the rolling CRPS test statistic between FQ-AL$_{250}$ and the three benchmark models with a window size of 500. Test statistics with particularly large magnitudes are capped for easier graphical representation. Positive values indicate favourable performance of FQ-AL$_{250}$ and values outside the red and blue area are significant at 5% and 1%. Figures based on weighted CRPS can be found in the supplementary materials.

Figure 6.9: FQ-AB$_{250}$: Uniformly weighted CRPS test statistic



We plot the rolling CRPS test statistic between FQ-AB$_{250}$ and the three benchmark models with a window size of 500. Test statistics with particularly large magnitudes are capped for easier graphical representation. Positive values indicate favourable performance of FQ-AB$_{250}$ and values outside the red and blue area are significant at 5% and 1%. Figures based on weighted CRPS can be found in the supplementary materials.

6.3.2 Multivariate Forecasting Accuracy

For the evaluation of multivariate forecasting accuracy we apply the energy score and the variogram score with $p = 0.5, 1, 2$. These values of p have been introduced by Scheuerer and Hamill (2015) and are considered typical choices (Jordan et al., 2017). Contrary to the CRPS results, the multivariate scoring rules encapsulate the accuracy for all eight marginals and their dependency into a single score which holistically quantifies the performance of the model on a given data set. Again, we start with a discussion of the MCS results on the entire out-of-sample period based on Tables 6.9 and 6.10, first for FQ-AL and subsequently for FQ-AB.

Overall, our Factor Quantile model FQ-AL maintains a good relative rank among all models, comparable to the more complicated GARCH models. In particular:

Exchange rate returns DCC-GARCH is represented in all superior sets but depending on the scoring rule, other models are included as well. Most notably, the variogram scores with $p = 0.5$ and $p = 1$ both retain FQ-AL$^{C}_{2000}$, making it the second best model. Further, the variogram score with $p = 2$ includes two Factor Quantile models with independent marginals and also CCC-GARCH. This is the only case where CCC-GARCH is included in the multivariate evaluation.

Interest rate changes All scoring rules strongly identify a single model as the best one but do not coincide in their choice. Variogram scores with $p = 0.5$ and $p = 1$ favour DCC-GARCH while the energy score and variogram score with $p = 2$ prefer FQ-AL$^{I}_{250}$ and FQ-AL$^{I}_{2000}$ respectively.

Commodity index returns The rankings by the four multivariate scores are largely consistent, ranking FQ-AL$^{C}_{250}$ as the best model. The variogram score with $p = 2$ deviates from this consensus slightly and prefers FQ-AL$^{C}_{2000}$ instead. Furthermore, the energy score additionally includes EDF$^{C}_{250}$ and EDF$^{I}_{250}$ in its superiors set.

Table 6.9: MCS p-values for FQ-AL: Multivariate scores

Model	$VS_{0.5}$	$VS_{1.0}$	$VS_{2.0}$	ES
Exchange rate returns				
CCC-GARCH	0.00	0.00	0.91**	0.00
DCC-GARCH	0.73**	0.76**	0.87**	1.00**
EDF^{C}_{250}	0.00	0.00	0.00	0.00
EDF^{C}_{2000}	0.00	0.00	0.00	0.00
EDF^{I}_{250}	0.00	0.00	0.00	0.00
EDF^{I}_{2000}	0.00	0.00	0.00	0.00
$FQ\text{-}AL^{C}_{250}$	0.19*	0.76**	0.14*	0.00
$FQ\text{-}AL^{C}_{2000}$	1.00**	1.00**	0.14*	0.00
$FQ\text{-}AL^{I}_{250}$	0.00	0.00	0.91**	0.00
$FQ\text{-}AL^{I}_{2000}$	0.00	0.00	1.00**	0.00
Interest rate changes				
CCC-GARCH	0.00	0.00	0.00	0.00
DCC-GARCH	1.00**	1.00**	0.00	0.00
EDF^{C}_{250}	0.00	0.00	0.00	0.00
EDF^{C}_{2000}	0.00	0.00	0.00	0.00
EDF^{I}_{250}	0.00	0.00	0.00	0.00
EDF^{I}_{2000}	0.00	0.00	0.00	0.00
$FQ\text{-}AL^{C}_{250}$	0.00	0.00	0.00	0.00
$FQ\text{-}AL^{C}_{2000}$	0.00	0.00	0.00	0.00
$FQ\text{-}AL^{I}_{250}$	0.00	0.00	0.14*	1.00**
$FQ\text{-}AL^{I}_{2000}$	0.00	0.00	1.00**	0.00
Commodity index returns				
CCC-GARCH	0.00	0.00	0.00	0.00
DCC-GARCH	0.00	0.00	0.00	0.00
EDF^{C}_{250}	0.00	0.00	0.00	0.26**
EDF^{C}_{2000}	0.00	0.00	0.00	0.00
EDF^{I}_{250}	0.00	0.00	0.00	0.99**
EDF^{I}_{2000}	0.00	0.00	0.00	0.00
$FQ\text{-}AL^{C}_{250}$	1.00**	1.00**	0.22*	1.00**
$FQ\text{-}AL^{C}_{2000}$	0.00	0.00	1.00**	0.00
$FQ\text{-}AL^{I}_{250}$	0.00	0.00	0.00	0.01
$FQ\text{-}AL^{I}_{2000}$	0.00	0.00	0.00	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

The results for FQ-AB are very similar to those of FQ-AL, which is expected since they share the same copula and also attained comparable performances during the assessment of the marginal forecasts. Minor differences include:

(i) The variogram score with $p = 2$ ranks FQ-AB$^C_{250}$ slightly higher than FQ-AL$^C_{250}$ for the commodity index returns;

(ii) In exchange rate returns, FQ-AB models are included in one more superior set than FQ-AL models, further closing the gap to DCC-GARCH.

Similar to the results in Section 6.3.1, accuracy of the models depends mostly on the data. Exchange rate returns are best explained with DCC-GARCH, commodity index returns with FQ-AL$^C_{250}$ or FQ-AB$^C_{250}$, and the best model in interest rates fluctuates between Factor Quantile specifications and DCC-GARCH depending on the choice of scoring rule. Both the energy and variogram score do not favour specific models predominantly and change their preferred model depending on the data. The only model that remains in all three superior sets for one scoring rule is FQ-AB$^C_{250}$.

Overall, both FQ-AL and FQ-AB stay in 75% of superior sets with at least one specification and the most successful versions FQ-AL$^C_{250}$ and FQ-AB$^C_{250}$ each remain in 33% and 67% of the sets. This is comparable to the 50% inclusion rate of DCC-GARCH and much stronger than CCC-GARCH and all individual EDF models which are in 8% of the superior sets at most.

Notably, Factor Quantile models again outperform all empirical distributions significantly, despite sharing the same calibration window and the same copula. This further shows that the noise reduction through our latent factor model improves the accuracy of the distribution forecast considerably.

The comparable performance of Factor Quantile models to DCC-GARCH, even with a simple Gaussian copula, is especially relevant since the latter is much more computationally intensive. As pointed out in Section 6.2, both Factor Quantile versions are at least 5 times faster and do not require additional attention to check for mis-calibrated parameters.

Table 6.10: MCS p-values for FQ-AB: Multivariate scores

Model	$VS_{0.5}$	$VS_{1.0}$	$VS_{2.0}$	ES
Exchange rate returns				
CCC-GARCH	0.00	0.00	0.09	0.00
DCC-GARCH	1.00**	0.40**	0.99**	1.00**
EDF^{C}_{250}	0.00	0.00	0.23*	0.00
EDF^{C}_{2000}	0.00	0.00	0.09	0.00
EDF^{I}_{250}	0.00	0.00	0.23*	0.00
EDF^{I}_{2000}	0.00	0.00	0.09	0.00
$FQ\text{-}AB^{C}_{250}$	0.00	1.00**	0.43**	0.57**
$FQ\text{-}AB^{C}_{2000}$	0.00	0.00	0.43**	0.00
$FQ\text{-}AB^{I}_{250}$	0.13	0.09	1.00**	0.01
$FQ\text{-}AB^{I}_{2000}$	0.13	0.00	0.09*	0.00
Interest rate changes				
CCC-GARCH	0.00	0.00	0.00	0.00
DCC-GARCH	1.00**	1.00**	0.00	0.00
EDF^{C}_{250}	0.00	0.00	0.00	0.00
EDF^{C}_{2000}	0.00	0.00	0.00	0.00
EDF^{I}_{250}	0.00	0.00	0.00	0.00
EDF^{I}_{2000}	0.00	0.00	0.00	0.00
$FQ\text{-}AB^{C}_{250}$	0.00	0.00	0.00	1.00**
$FQ\text{-}AB^{C}_{2000}$	0.00	0.00	0.00	0.01
$FQ\text{-}AB^{I}_{250}$	0.00	0.00	0.16*	0.00
$FQ\text{-}AB^{I}_{2000}$	0.00	0.00	1.00**	0.00
Commodity index returns				
CCC-GARCH	0.00	0.00	0.00	0.00
DCC-GARCH	0.00	0.00	0.00	0.00
EDF^{C}_{250}	0.00	0.00	0.00	0.00
EDF^{C}_{2000}	0.00	0.00	0.00	0.00
EDF^{I}_{250}	0.00	0.00	0.00	0.00
EDF^{I}_{2000}	0.00	0.00	0.00	0.00
$FQ\text{-}AB^{C}_{250}$	1.00**	1.00**	0.28**	1.00**
$FQ\text{-}AB^{C}_{2000}$	0.00	0.00	1.00**	0.00
$FQ\text{-}AB^{I}_{250}$	0.00	0.00	0.00	0.00
$FQ\text{-}AB^{I}_{2000}$	0.00	0.00	0.00	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

There are several distinct features of the multivariate results, especially in comparison to the univariate evaluation in Section 6.3.1:

(i) Contrary to the univariate analysis, models with longer calibration windows perform better and are now present in the superiors sets. This is likely explained by the improved correlation matrix since an estimation based on a larger sample size reduces the standard errors.

(ii) The performance of DCC-GARCH is much better in the multivariate comparison than in the prior univariate one, even for exchange rate returns where CCC-GARCH was included in more superior sets than DCC-GARCH. This suggests that the time-varying conditional correlation structure is an improvement over the constant conditional correlation that requires strong assumptions not fulfilled for many assets (Tsui and Yu, 1999).

(iii) More generally, the univariate performance does not seem to influence the multivariate scoring rules significantly. This lack of sensitivity is particularly notable for the interest rate changes. Factor Quantile models dominated the superior sets based on all weights for CRPS but this is not reflected in the superior sets of the multivariate scoring rules. Only the energy score and the variogram score with $p = 2$ prefer the Factor Quantile models in both the FQ-AL and the FQ-AB comparison, indicating that the other two parameterisations of the variogram score place less importance on the marginal performances and more on the dependency structure. This further emphasizes the recommendations of Gneiting et al. (2008) and Scheuerer and Hamill (2015) that multivariate evaluation should include both univariate and multivariate scoring rules.

Before we discuss the general applicability of multivariate scoring rules, we examine the rankings over time. Table 6.11 shows how many of the four multivariate scoring rules include each particular model in the superior set. The tables with the underlying p-values for each scoring rule can be found in the supplementary materials.

As in Section 6.3.1, we consider three sub-periods ranging from (1) March 2007 – June 2010, (2) June 2010 – June 2014 and (3) June 2014 – June 2018 respectively.

Table 6.11: Comparison of multivariate performance over time

Model	Exchange rates				Interest rates				Commodity indices			
	(∗)	(1)	(2)	(3)	(∗)	(1)	(2)	(3)	(∗)	(1)	(2)	(3)
Factor Quantile with last principal components (FQ-AL)												
CCC-GARCH	1	1	0	0	0	0	1	1	0	1	0	0
DCC-GARCH	4	4	2	0	2	2	2	3	0	0	0	1
EDF^{C}_{250}	0	0	0	0	0	0	0	0	1	0	1	4
EDF^{C}_{2000}	0	3	0	0	0	0	0	0	0	3	0	1
EDF^{I}_{250}	0	0	0	0	0	0	1	0	1	0	1	4
EDF^{I}_{2000}	0	3	0	0	0	0	0	0	0	3	0	1
FQ-AL^{C}_{250}	1	1	2	4	0	0	1	1	3	3	4	3
FQ-AL^{C}_{2000}	2	3	0	0	0	0	0	0	1	2	0	4
FQ-AL^{I}_{250}	1	0	0	0	1	2	0	1	0	0	0	0
FQ-AL^{I}_{2000}	1	0	1	0	1	0	1	2	0	0	0	0
Factor Quantile with asymptotic bagging (FQ-AB)												
CCC-GARCH	0	0	0	0	0	0	1	0	0	0	0	0
DCC-GARCH	4	4	3	0	2	2	2	2	0	0	0	0
EDF^{C}_{250}	0	0	1	0	0	0	0	0	0	0	2	0
EDF^{C}_{2000}	0	1	0	0	0	0	0	0	0	0	0	0
EDF^{I}_{250}	0	0	1	2	0	0	0	0	0	0	2	0
EDF^{I}_{2000}	0	1	0	0	0	0	0	0	0	0	0	0
FQ-AB^{C}_{250}	3	0	2	2	1	0	1	1	4	1	3	4
FQ-AB^{C}_{2000}	1	3	0	0	0	0	0	0	1	3	0	0
FQ-AB^{I}_{250}	1	0	0	0	0	2	0	0	0	0	0	0
FQ-AB^{I}_{2000}	0	0	1	1	1	0	1	1	0	0	0	0

This table lists the number of times each model is included in one of the superior sets with $\alpha = 0.25$ for the multivariate scores. Since we consider 4 different scoring rules, each model can be included at most 4 times. Column (∗) uses the entire out-of-sample periods while columns (1), (2) and (3) are restricted to the sub-periods March 2007 – June 2010, June 2010 – June 2014 and June 2014 – June 2018 respectively.

The general ranking remains similar in each sub-period with DCC-GARCH, FQ-AL^{C}_{250} and FQ-AB^{C}_{250} as clearly the most highly ranked models. Over all data sets and sub-periods, FQ-AL^{C}_{250} remains in 53% of the superior sets compared to 39% of DCC-GARCH and 19% of EDF^{C}_{2000} or EDF^{I}_{2000} which are the best performing benchmark models. The FQ-AB comparison yields a similar outcome with an inclusion rate of 39% for FQ-AB^{C}_{250}, versus 36% of DCC-GARCH and 14% of EDF^{I}_{250}.

Within the sub-periods, the empirical distribution functions are included in more superior sets than they are when using the entire sample period. Several other

models are also included in specific periods but do not remain in the superior set consistently. This is likely a combination of two effects:

(i) The models fit the changing data generating process better in certain periods. For instance, this seems to be the case for FQ-AL$^C_{250}$ in exchange rate returns where it replaces DCC-GARCH in the latter sub-periods as the best model.

(ii) The drastically reduced sample size increases the uncertainty of MCS which leads to more inclusions in general. This is likely the case where more models than usual are included in the superior sets as in sub-period (1) for exchange rate returns and sub-period (3) for commodity index returns.

Furthermore, the multivariate scoring rules are more irregular than their univariate counterparts. For instance, FQ-AB$^C_{250}$ is included in the superior set of 3 scoring rules in exchange rate returns given the entire sample but performs weakly in each sub-period. In contrast, none of the scoring rules prefer EDF$^C_{250}$ overall although it performs well in some sub-periods. These patterns occur since the aggregation of inclusions over sub-periods does not consider the performance holistically. This further motivates the use of long out-of-sample periods for accuracy evaluation.

Our study additionally highlights issues with multivariate scoring rules that arise due to the high degrees of freedom. In some cases, the resulting ranking varies depending on the choice of scoring rule. This is particularly evident in the interest rate data set where it remains unclear which model actually performs best since the four scoring rules each identify one of three models as the best one. For exchange rate and commodity index returns, the problem is also present, but less severe. Some models are only included for specific scoring rules but there is an overall consensus for the superior performance of one model. Additionally, some scoring rules favour models with independent marginals. These models certainly do not capture the dependency between the assets accurately and therefore we should not expect them to be included in the superior sets. Either the dependency structure of the Gaussian copula and the GARCH models is not suitable for the data, leading the independent versions to be relatively good despite their shortcomings, or the scoring rules fail to

identify the actual best model. This happens particularly often with the variogram score with $p = 2$ which retains 12 independent models over both MCS comparisons and all sub-periods. The energy score follows with the inclusion of 8 independent models. Only 6 independent models remain in the superior set for variogram score with $p = 0.5$ and $p = 1$ each. We analyse those issues further in Chapter 7 where we design a simulation study and test the ability of each multivariate scoring rule to identify the best model.

Chapter Seven

Simulation Study

7.1 Simulation Design 153
7.2 Simulation Results 157
7.2.1 Sample Mean Comparison 159
7.2.2 Error Rate of Scoring Rules 162
7.2.3 Discrimination Heuristic of Scoring Rules 166

Our discussion in Section 5.2 introduced several univariate and multivariate proper scoring rules that are prevalent in the literature. While it is agreed upon that these offer a sound way to quantify the accuracy of probabilistic forecasts (Winkler, 1996; Gneiting and Ranjan, 2011), the question of which score to use remains largely open (Gneiting and Raftery, 2007). Conventional wisdom dictates to apply a suitable scoring rule for the application at hand (Machete, 2013) but this only provides a few requirements and does not sufficiently restrict the selection.[1]

The choice of the scoring rule is much less of a problem in the univariate case which is likely the reason why this issue has not been rigorously addressed yet. Although there might be slight deviations, the rankings that univariate scoring rules provide mostly coincide, so that there are no conflicting conclusions (Staël von Holstein, 1970; Winkler, 1971; Bickel, 2007). Therefore, in most settings any scoring rule may be applied.

Unfortunately, the same does not hold true for multivariate scoring rules. Our empirical study in Chapter 6 clearly demonstrates that the energy score and different parameterisations of the variogram score rules do not generally recommend the same distribution forecast. The high degree of freedom leads to a loss of information during the encapsulation into a single score and forces the multivariate scoring rules to focus on different aspects of goodness that may be contradictory. This begs the question of which score to trust and, more broadly, if a single score can adequately reflect the entirety of the relevant information in higher dimensions at all.

The primary goal of a scoring rule is to provide a correct ranking of models. This is in part covered by propriety since the true model receives the lowest attainable score. In practical applications, however, there are further considerations that are of relevance:

(i) Propriety concerns only the expectation. Given a sample mean based on a realistic sample size, even strictly proper scoring rules may lead to wrong

[1] As pointed out in Section 5.2, scoring rules have varying assumptions for propriety and compare different forecasting types, e.g. density forecast, distribution forecasts or ensemble forecast, that are sometimes not easily interchangeable.

inferences. Pinson and Tastu (2013) quantify this likelihood heuristically through their discrimination heuristic which measures the distance between the scores of competing models. A relatively large distance may be interpreted as a sign of robust rankings.

(ii) Generally, the true distribution is not known and none of the models in the comparison may accurately reflect the true distribution (Elliott and Timmermann, 2008). Typically only misspecified models are compared against each other, with no guarantee that a 'better' forecast receives the lower score.[2] We contend that this is difficult to avoid without strict definitions of goodness, possibly through utility functions.

For instance, strictly proper scoring rules with low discriminatory power may assign very similar scores to competing models, so that the score expectation of the true distribution is only slightly below that of misspecified models. Since the true distribution receives the lowest expectation, the scoring rule is strictly proper but the small difference between the score expectations may not be captured by the sample means in empirical applications which leads to erroneous rankings of the competing models. This motivates additional requirements beyond propriety for scoring rules which quantify their sensitivity.

Despite the critical practical implications regarding the choice of the multivariate scoring rule, very little formal research has been conducted so far. This may be attributed to the overall small literature on multivariate forecasting evaluation paired with the difficulty to evaluate scoring rules without strong assumptions about the specific setting or data.

As mentioned in our literature review in Chapter 2, there are several studies that analyse proper scoring rules analytically but they do not yield sufficient guidance on the scoring rule selection apart from some generic suggestions or are limited to a binary setting (Buja et al., 2005; Merkle and Steyvers, 2013; Johnstone et al., 2011).

[2]In this case, proper scoring rules still enforce honest forecasts since forecasters maximize their expected score by volunteering their true beliefs.

In contrast, studies of multivariate proper scoring rules have mostly been limited to simple simulations settings and are discussed by Pinson and Tastu (2013) and Scheuerer and Hamill (2015). However, most of these studies consider various elliptical and light-tailed Gaussian distributions which do not reflect realistic conditions in finance or economics adequately. In fact, in the only case where the data generating process (DGP) is not a Gaussian but a Poisson distribution, the results varied strongly from the Gaussian setting. All scores but the variogram score with $p = 0.5$ had ranking issues and may identify the wrong model as the correct one (Scheuerer and Hamill, 2015). Furthermore, some of the findings may be attributed directly to the simulation design. For instance, Pinson and Tastu (2013) assume a bivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = (\mu, \mu), \qquad \boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

as DGP and impose misspecifications in mean, variance and covariance by changing the correct parameters to

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}, \hat{\mu}), \qquad \hat{\boldsymbol{\Sigma}}_{\hat{\sigma}^2} = \hat{\sigma}^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \hat{\boldsymbol{\Sigma}}_{\hat{\rho}} = \sigma^2 \begin{pmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{pmatrix}$$

respectively. Hence, misspecifications in mean and variance affect both dimensions in the simulation which may yield an unfair comparison to the correlation. In fact, a deviation in variance affects all elements of the covariance matrix while a deviation in correlation is restricted to changes in the anti-diagonal elements. Therefore, direct comparisons between the resulting changes in the energy score may be difficult, particularly since the misspecified parameters were chosen arbitrarily around the correct parameters. Also, the sensitivity to individual changes does not reveal how the scoring rules react to a combination of misspecifications which is likely to be the case in practical applications. Because everything is encapsulated into a single score, simultaneous changes may cancel each other out or augment each other.

During our simulation study, we generalize the discrimination heuristic of Pinson and Tastu (2013) and analyse the ability of different scoring rules to identify the

true distribution. We extend previous studies in the literature by choosing a realistic simulation setting that better approximates the conditions in practical applications. This is reflected in our simulation design which employs daily USD-denominated exchange rates from 1999 – 2018; US interest rates from 1994 – 2018; and Bloomberg investable commodity indices from 1991 – 2018 together with popular models that are regularly used within those data sets. As mentioned in our literature review in Chapter 2, prior studies only considered various parametric distributions as DGP and misspecified models. Section 7.1 describes the design of the simulation study and motivates the choices we make. All results are discussed in Section 7.2. For reasons of space, some tables and figures are only available electronically in the supplementary materials.

7.1 Simulation Design

Our simulation study quantifies the ability of the energy score and the variogram score with $p = 0.5, 1, 2$ to distinguish the correct DPG from misspecified models. These values of p have been used by Scheuerer and Hamill (2015) and are considered typical choices (Jordan et al., 2017). As mentioned in Section 5.2, we exclude multivariate scoring rules which require a density forecast because our forecasting models produce ensemble forecasts. Further, we do not consider the Dawid-Sebastiani score because it only relies on the mean and the covariance of the forecasts and cannot distinguish predictions with differences only in higher moments which is often detrimental in financial and economic data sets.

Contrary to other papers in the literature, we use a realistic setting with actual data and three types of distribution forecasting methods. Differences between the DGP and misspecified models generally occur in multiple ways, unlike the ceteris paribus examinations of Pinson and Tastu (2013) and Scheuerer and Hamill (2015). Further, we generalize the approach of Pinson and Tastu (2013) to compare the discrimination ability of several scoring rules and introduce the error rate as an additional heuristic for the sensitivity of scoring rules.

In our simulation setting, we control the DGP such that at each time t we know the true distribution. We apply the same models as in our empirical study in Chapter 6 which are summarized in Table 6.3 on systems of daily, eight-dimensional USD-denominated exchange rates, interest rates and Bloomberg investable commodity indices that we discussed in Section 6.1. For simplicity, we only consider models that incorporate the dependency structure, so we end up with eight competing models in total with the same parametrisation as in the previous chapters: (i) Two FQ-AL models, (ii) two FQ-AB models, (iii) two EDF models and (iv) two multivariate GARCH models. Therefore, each model has one associated model that is similar but differs either in the calibration length or the correlation structure. To reduce the dependence of our simulation study on a specific DGP, we repeat the simulation eight times and rotate the choice of DGP across all models.

The simulation for a specific DGP uses observations up to time T. We compare the ability of the multivariate scoring rules to distinguish the DGP from the misspecified models based on their distribution forecasts for $T + 1$:

Stage 1 Given historical data up to a time T, we calibrate all models including our choice for the DGP. Then, we forecast distributions for $T + 1$.

Stage 2 We draw 5,000 samples from the forecasted distribution of the DGP and assume each sample is a realisation at time $T + 1$.

Stage 3 For each of the 5,000 realisations we quantify the performance of all models. That is, we draw an ensemble of 10,000 samples from each distribution forecast and calculate the scores based on the current realisation.[3] This leaves us with 5,000 scores for each of the models for our subsequent analysis.

Depending on the realisation, the scoring rules may favour a model other than the DGP but the sample mean based on all 5,000 scores should be the smallest for the DGP. This, of course, is because the distribution of the DGP is used to generate the realisations. A good scoring rule should assign the lowest scores to the DGP and also produce robust rankings over the entire evaluation period. As Pinson and Tastu (2013) point out, a large distance between the scores of the DGP and alternative models may help to avoid erroneous conclusions.

We evaluate the scoring rules at the first date of each quarter in our evaluation period which yields simulations on 50, 66, and 82 dates in USD-denominated exchange rates, US interest rates and Bloomberg investable commodity indices with eight dimensions respectively. Since we have eight possible DGPs, this leads to approximately 1,600 applications of the simulation above for each of the four multivariate scoring rules. This setting gives us a very detailed view on the discrimination ability for each scoring rule over time and for various choices of the DGP.

Our simulation design reflects optimistic conditions for the scoring rules since it knows the distribution of the DGP and samples a very large number of realisations at

[3]It is easy to sample with our forecasting models. Hence, we apply a large ensemble that outlines the distribution forecasts in detail.

each time t. In practice, we only observe one realisation and therefore must consider the scores over a large period instead. Hence, each simulation at time t corresponds to an evaluation with the scoring rules based on 5,000 out-of-sample observations where the underlying DGP is stationary. We also compare the scoring rules on smaller sub-samples with only 100 realisations to approximate a more realistic setting, in which the length of the out-of-sample period is restricted due to lack of data.

It is important to note that the performance of the models in this chapter is not reflective of their actual forecasting accuracy. All models use historical information to forecast their joint distribution but are then evaluated against samples from the chosen DGP rather than the realisation of the original time series. The models are therefore punished if their forecast deviates from that of the DGP. However, this simulation design allows us to assess the performance of the scoring rules. Within the competing models is the DGP and a good scoring rule must be able to distinguish other alternative models from the true distribution. We refer to Chapter 6 for a discussion on the relative performances of each model.

In this simulation setting, we know the true DGP at each forecasting date and avoid issues with re-calibration. An alternative simulation design for which the distribution at time $T + 1$ also is known would be to create an artificial time series using the DGP and then re-estimating all models on this time series. This approach is chosen by both Pinson and Tastu (2013) and Scheuerer and Hamill (2015). We opt against this design for two main reasons:

(i) The re-calibration of the DGP on a time series produced by itself does not necessarily yield a good fit. This issue is especially relevant for the more complicated GARCH models where estimation errors are expected to be non-neglectable. Therefore, the DGP may produce a different distribution forecast after re-calibration and we could not ensure that the DGP obtains the lowest score. Simpler misspecified models might be better than the correct model with wrong parameters (Elliott and Timmermann, 2016).

(ii) It is unclear how parameters of the DGP should be chosen in the alternative simulation. In contrast, our model parameters in the simulation study reflect realistic market conditions.[4]

[4]Choosing specific parameters ex-ante can be avoided by calibrating the model on the data sets but this would require calibrations on real data, similar to our approach.

7.2 Simulation Results

We analyse the energy score and three parameterisations of the variogram score with respect to their ability to identify the DGP under the simulation design discussed in Section 7.1. The scores assigned by scoring rule s to model m at time t with model m^* as DGP are defined as $\mathbf{S}_t^s(m, m^*)$ where $m, m^* = 1, \ldots, M$. With N realisations at each time t,

$$\mathbf{S}_t^s(m, m^*) = \left(S_{1,t}^s(m, m^*), \ldots, S_{N,t}^s(m, m^*)\right)'.$$

As discussed in Section 7.1, we use $N = 5000$ and $M = 8$. To examine the discrimination ability of each scoring rule at time t, we apply the entire sample of 5,000 scores but also smaller sub-samples with 100 scores that reflect more realistic conditions. These correspond to an out-of-sample evaluation with 5,000 and 100 observations.

It should again be emphasized that the performance of our eight forecasting models in this section is not an indicator for their real relative accuracy. Because we impose a distribution through the DGP, the scores rather measure the closeness of the distribution forecasts to the distribution of the DGP.

We begin our discussion with the sample score mean, focussing on exchange rate returns and a DCC-GARCH as DGP in Section 7.2.1. Then, we generalize these results to multiple DGPs and data sets by introducing the error rate in Section 7.2.2 as the percentage of cases in which a misspecified model receives a lower (i.e. better) score than the DGP. Further, we analyse the deviation between the scores of misspecified models and that of the DGP for each scoring rule. We illustrate the distribution of these deviations and generalize the discrimination heuristic proposed by Pinson and Tastu (2013) in Section 7.2.3, in order to compare multiple scoring rules.

Our simulation study on three data sets with eight DGP shows that the variogram scores with $p = 0.5$ and $p = 1$ have a lower error rate than the energy score or the variogram score with $p = 2$. Further, the discrimination heuristic indicates

consistently large distances between the sample score means of the DGP and those of misspecified models for the variogram score with $p = 1$. These results are robust for all choices of the DGP and data set. Hence, our findings identify the variogram scores with $p = 0.5$ or $p = 1$ as the best scores overall. Simultaneously, the simulation study shows that wrong rankings can be frequent, especially with smaller sample sizes. This issue becomes even more relevant in practical applications, where we additionally encounter other problems such as calibration errors which further complicate the identification of the DGP. We therefore suggest the use of multiple types of scoring rules for the evaluation in higher dimensions.

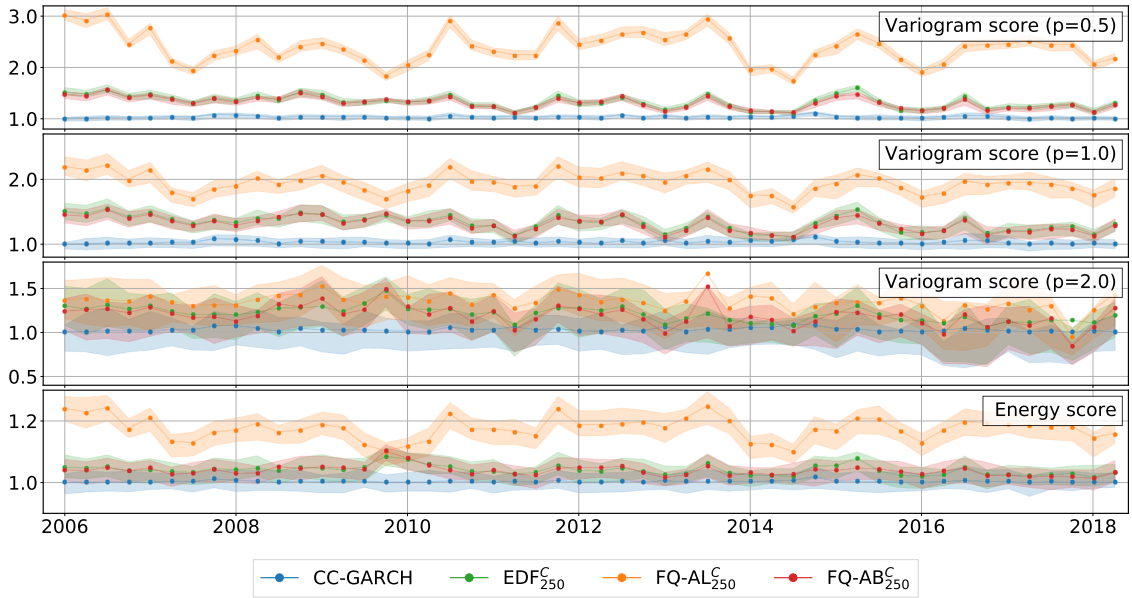
7.2.1 Sample Mean Comparison

We begin the analysis of the multivariate scoring rules with a comparison of their sample score mean for each model. Figure 7.1 uses a DCC-GARCH as DGP for exchange rate returns and shows the ratio

$$\frac{1}{N}\sum_{i=1}^{N}\left(S_{it}^{s}\left(m,m^{*}\right)/S_{it}^{s}\left(m^{*},m^{*}\right)\right) \tag{7.1}$$

for four selected models. This corresponds to the relative distance between the sample score mean of the DGP and that of misspecified models. The shaded areas cover everything between the 0.25- and 0.75-quantiles for the sample mean based on a sample size of 100 instead of 5,000. These confidence intervals are generated through a statistical bootstrap with 5,000 repetitions. We limit the illustration to four models only for clarity, but the results are comparable when other misspecified models, DGPs or data sets are considered. Figures for other DGPs and data sets can be found in the supplementary materials.

Figure 7.1: Average scores relative to score of DGP (USD exchange rates)



The figure illustrates the relative sample score mean in Equation 7.1 based on 5,000 scores. A value larger than 1 means that the scoring rule is on average able to distinguish between the misspecified model and DCC-GARCH to identify the true DGP. We generate a confidence interval covering the area between the 0.25- and 0.75-quantiles of the sample mean based on a sample size of 100 through bootstrap with 5,000 repetitions.

The results based on the sample mean of 5,000 scores indicate that all four scoring rules manage to evaluate the models successfully. Due to propriety, they assign the lowest expectation to the DGP which is why almost none of the sample score mean fall below 1 in Figure 7.1. Further, CCC-GARCH generally obtains the lowest score among all misspecified models which is expected given its similarity to DCC-GARCH. The scores can distinguish distributions which differ only in their marginals and are able to identify FQ-AL$^C_{250}$ as one of the misspecified models with great confidence.[5] In contrast, the difference between FQ-AB$^C_{250}$ and EDF$^C_{250}$ is less pronounced which means that they produce predictions of equal similarity to the distribution forecast of DCC-GARCH.

Both the variogram score with $p = 0.5$ and $p = 1$ show clear and robust rankings between the misspecified models and distinguish them from the DGP. The discrimination ability is weaker for the energy score. As pointed out by Pinson and Tastu (2013) and Scheuerer and Hamill (2015), the energy score changes only by a small amount between the DGP and other models. This is evident in Figure 7.1 as well, where the average score of the worst model is only 25% larger than that of the DGP. In comparison, the variogram scores with $p = 0.5$ and $p = 1$ assign average scores over 200% and 100% larger than that of the DGP respectively. Unlike the other scoring rules, the variogram score with $p = 2$ changes the rankings at several times and is also the only scoring rule which makes wrong inferences even with a large sample size of 5,000 scores. For instance, FQ-AB$^C_{250}$ is preferred over the DGP around the end of 2017. Hence, the energy score and variogram scores with $p = 0.5$ and $p = 1$ may be preferable to the variogram score with $p = 2$.

However, there are vast differences in the discrimination ability which can lead to wrong inferences in smaller sample sizes:

(i) Despite the overall success of the variogram score with $p = 0.5$ and $p = 1$, wrong inferences may occur with only 100 samples. The shaded areas of

[5]The number of latent factors in FQ-AL$^C_{250}$ produces a much sharper forecast than that of alternative models. As shown in Chapter 6, this yields good forecasts. However, due to the narrow range of the predictions, FQ-AL$^C_{250}$ is an easily identifiable model in this simulation study. All our results in this section persist if we exclude the FQ-AL models from the analysis.

CCC-GARCH dip below 1 frequently which means that a slightly misspecified model may be chosen over the DGP.

(ii) This is also true for the energy score but to a much larger extent. Besides CCC-GARCH, FQ-AB$^C_{250}$ and EDF$^C_{250}$ are also assigned lower scores than the DGP in 2010, 2013, 2016 and 2017. Overall though, the energy score still manages to produce a clear ranking that is mostly accurate.

(iii) The variogram score with $p = 2$ largely fails to yield any meaningful results with the smaller sample size. The rankings can change considerably, and all models obtain a lower sample mean than the DGP at various times. Even FQ-AL$^C_{250}$, which is regarded as the worst model by all other scoring rules, has lower scores than DCC-GARCH around 2016. Additionally, the variogram score with $p = 2$ may assign scores of very large magnitude that greatly affect the sample mean. This is visible in Figure 7.1 in two aspects: (i) The scoring rule has wide confidence intervals and (ii) the sample mean is at times higher than the sample 0.75-quantile. This is, for instance, the case around the end of 2013.

These initial findings suggest that variogram score with $p = 0.5$ and $p = 1$ offer superior discrimination ability to the more popular energy score. The variogram score with $p = 2$ performs very poorly and may yield erroneous rankings of the forecasting models, even with a very large sample of scores.

7.2.2 Error Rate of Scoring Rules

The sample score means clearly indicate that scoring rules may yield erroneous rankings in smaller samples. For some realisations, the lowest score may be assigned to a model that is not the DGP. We study this probability in our simulation study by introducing an error rate measure for each scoring rule and by analysing the distribution of

$$\mathbf{S}_t^s\left(m, m^*\right) - \mathbf{S}_t^s\left(m^*, m^*\right), \tag{7.2}$$

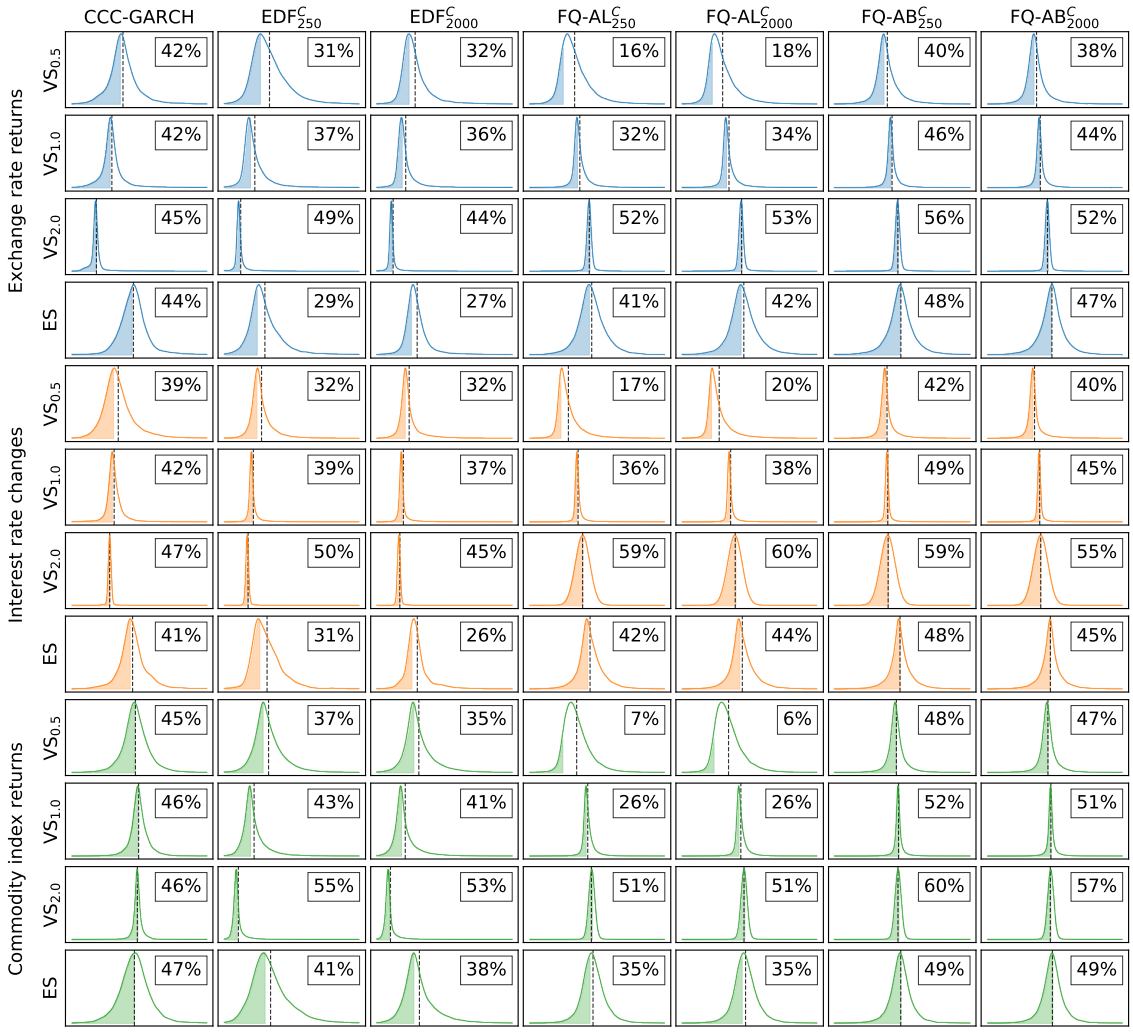
which constitutes the absolute differences between the scores assigned to each model in comparison to the scores of the DGP.

Figure 7.2 shows the results with DCC-GARCH as DGP and uses the scores for all t to generate the density. Each column of the figure illustrates the density of Equation 7.2 for a specific misspecified model, under various scoring rules and data sets. We include the error rate in the upper right corner of each sub-figure which shows the probability that Equation 7.2 yields a negative value. For clarity, we do not use the same x-axis for all sub-figures but show all values between the 0.001- and 0.999-quantiles of each distribution. This means that the magnitude of the error is not visible in these figures but instead we gain insight on the shape of the error density. Figures for alternative DGPs can be found in the supplementary materials.

Overall, Figure 7.2 shows that the probability of getting scores which are lower than that of the DGP is high and varies around 31% and 54%, depending on the data set and scoring rule. The variogram score with $p = 2$ particularly often assigns lower scores to misspecified models. This happens in 50%, 54%, 53% of cases for exchange rate returns, interest rate changes and commodity rate returns respectively and is therefore around 60% worse than the error rate of the variogram score with $p = 0.5$. This scoring rule achieves the lowest error rate, followed by the variogram score with $p = 1$ and the energy score.

It is important to note that the error rate is only a binary statistic which does not take into account the magnitude by which the scores of misspecified models

Figure 7.2: Density of differences between scores with DCC-GARCH as DGP

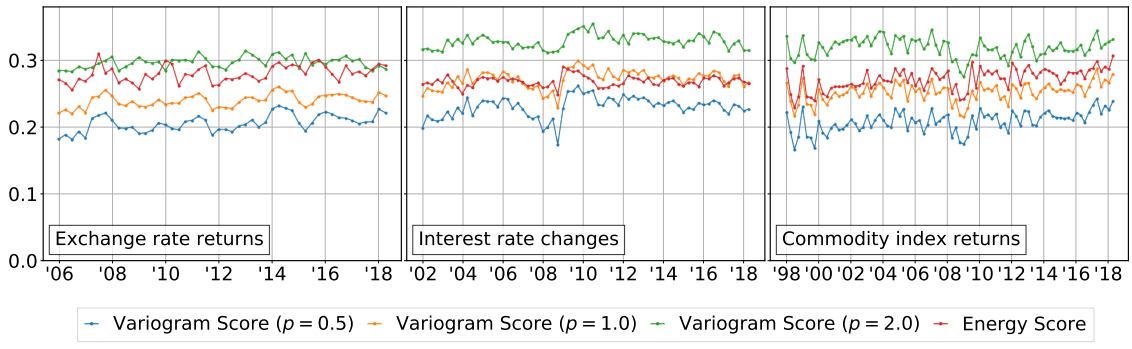


This figure displays the density of the difference between the scores of the DGP and the misspecified models described in Equation 7.2. A Gaussian kernel is used to smooth the densities. The shaded areas correspond to negative values, where a lower score is assigned to the misspecified models. In the upper right corner of each sub-figures, the probability of the shaded area is displayed. The dotted vertical line shows the expectation of the density. For clarity, we limit the sub-figures to values between the 0.001- and 0.999-quantiles. Figures for alternative DGPs can be found in the supplementary materials.

are smaller than that of the DGP. By averaging over a sample of scores, the error rate decreases, until it reaches zero due to the propriety of the scoring rules. The number of samples needed for a sample mean that favours the DGP depends on the shape of the distribution. If the tail of the shaded area is small in comparison to the tail of the non-shaded area, a small sample might be sufficient. However, many of the distributions in Figure 7.2 are approximately symmetric which means that large positive and negative values in Equation 7.2 are equally likely. As an additional indicator for the convergence speed, we illustrate the expectation of the distributions with a dotted line. These are always non-negative due to propriety of the scoring rules, but an expectation far right from the shaded area corresponds to a faster convergence towards lower sample score means for the DGP. Again, the values are generally close to the cut-off point 0 which suggests slow convergence towards positive sample mean scores.

The average error rate over all DGPs for the evaluation period of the multivariate scoring rules is compared in Figure 7.3. Similar to Figure 7.2, we examine the number of times the score of a misspecified model is lower than that of the DGP but now consider the error rate across multiple choices of the DGP.

Figure 7.3: Error rates of scoring rules



The error rates show how often a misspecified model is assigned a lower score than the DGP. Higher values are associated with inferior scoring rules and more frequently wrong inferences.

The results of Figure 7.3 are similar to Figure 7.2. The variogram score with $p = 2$ has a significantly higher error rate that is more than 47% higher than that of the variogram score with $p = 0.5$. This is also consistent with Figure 7.1 where

misspecified models were preferred over the DGP. Again, there is a clear ranking of the scoring rules that persists with all three data sets and the entire evaluation period. For the variogram scores, the error rate increases with the parameter p and the error rate of the energy score typically falls between the variogram score with $p = 1$ and $p = 2$.

7.2.3 Discrimination Heuristic of Scoring Rules

As an additional measure for the discrimination ability, we consider a simple heuristic that examines the relative distance of the scores between the models. A large distance may indicate that the ranking of the scores is reliable and not prone to change depending on the sample size. This approach has been suggested by Pinson and Tastu (2013) who compare the sensitivity of the energy score to various misspecifications. Given

$$\overline{\mathbf{S}}_t^s\left(m, m^*\right) \coloneqq \frac{1}{N} \sum_{i=1}^{N} S_{it}^s\left(m, m^*\right),$$

they utilize a Gaussian DGP and measure the sensitivity pairwise through

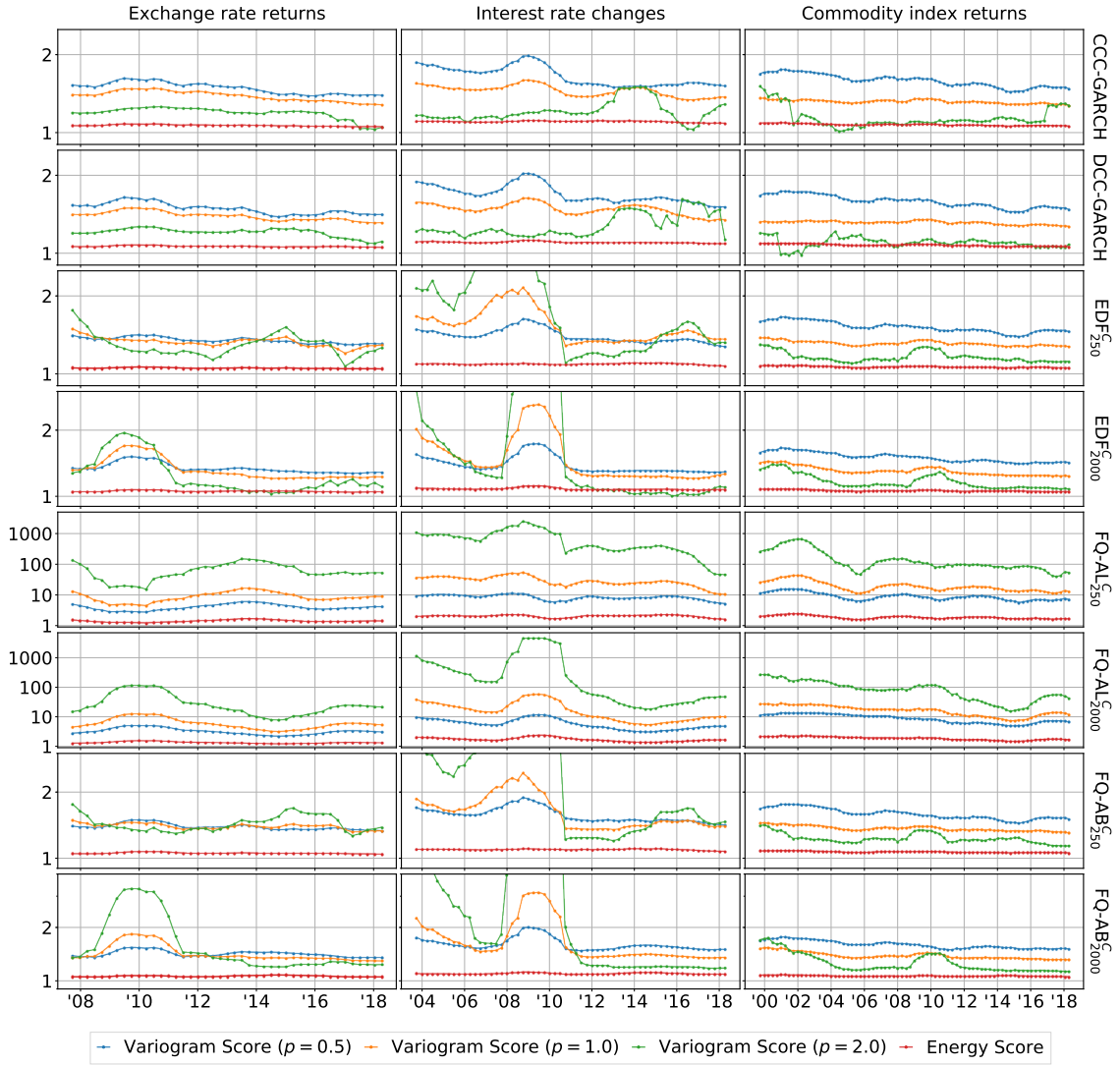
$$\frac{\overline{\mathbf{S}}_t^s\left(m, m^*\right) - \overline{\mathbf{S}}_t^s\left(m^*, m^*\right)}{\overline{\mathbf{S}}_t^s\left(m^*, m^*\right)}$$

with misspecified models that deviate only in one aspect (e.g. mean, variance or correlation) from the DGP. We adjust their measure to consider the discrimination across scoring rules over multiple misspecified models. To this end, we propose a generalized discrimination heuristic that is defined as

$$d_t^s(m^*) = \frac{1}{M} \sum_{m=1}^{M} \frac{\overline{\mathbf{S}}_t^s\left(m, m^*\right)}{\overline{\mathbf{S}}_t^s\left(m^*, m^*\right)}. \tag{7.3}$$

Through the consideration of multiple models, we go beyond ceteris paribus sensitivities to obtain more general results. Our misspecified models combine various misspecifications at once and are therefore more similar to the settings under which the proper scoring rules are applied in practice. We do not subtract the scores of the DGP from those of the misspecified models in the numerator, but this does not affect the rankings of the scoring rules regarding their values of the discrimination heuristic. Our adjusted discrimination heuristic is depicted in Figure 7.4 with a logarithmic scale. Contrary to Figure 7.1, the heuristic summarizes the results of multiple DGPs and all three data sets.

Figure 7.4: Discrimination heuristic of scoring rules



We display the discrimination heuristic of Equation 7.3 for all three data sets and eight DGPs with a logarithmic scale. Scoring rules which separate the scores of misspecified models and the DGP by a larger relative distance are assigned higher values for the discrimination heuristic. We smooth the discrimination heuristic with a moving average of 8 observations to improve the interpretability of the figure, but the same patterns are present in case no smoothing is applied.

Figure 7.4 shows a clear distinction between the models with similar results in all scenarios, but the preferences of the discrimination heuristic vary slightly depending on the data set and the DGP. Overall, there are several distinct features:

(i) The energy score is always the scoring rule with the lowest discrimination heuristic. This, again, is in accord with prior simulation studies by Pinson and Tastu (2013) and Scheuerer and Hamill (2015). Across all data sets and DGPs, the energy score only receives an average discrimination heuristic of 1.23, compared to 2.79, 5.30 and 78.13 for the variogram score with $p = 0.5$, $p = 1$ and $p = 2$.

(ii) In all cases, the variogram score with $p = 1$ is the scoring rule with the second highest discrimination heuristic.

(iii) The variogram score with $p = 2$ achieves in some settings extremely high values for the discrimination heuristic, but is also the only scoring rule which receives values below 1. This occurs in commodity index returns with DCC-GARCH as DGP. For those t, the model ranking of the variogram score with $p = 2$ is erroneous and multiple misspecified models receive lower scores than the DGP;

(iv) The scoring rule with the highest discrimination heuristic varies depending on the choice of data and DGP but exhibits a pattern. In most cases, the variogram score with $p = 0.5$ has the highest discrimination heuristic, but it is surpassed by the variogram score with $p = 2$ during some periods and when FQ-AL models are used as DGP.

The high discrimination heuristic of some variogram scores with $p = 2$, despite the poor performance in Figure 7.1 can be explained by Figure 5.2 and our discussion on the effect of different choices of p in Section 5.2.3. Generally, the variogram score with $p = 2$ outputs a large range of scores, some of which may be vastly larger in magnitude than others. These outliers shift the sample mean in Figure 7.1 to a larger value than the sample 0.75-quantile and also affect the discrimination heuristic to a similar extent. For instance, in exchange rate returns with DCC-GARCH as DGP,

the largest summand of Equation 7.3 takes a value around 4,700. In comparison, the largest summand of the energy score, variogram score with $p = 0.5$ and $p = 1$ are 17, 76 and 141 respectively.

The second power in the formula of the variogram score with $p = 2$ further amplifies large distances between models. Therefore, the variogram score with $p = 2$ achieves a particularly high discrimination heuristic when the models are easily distinguishable. The cases where the variogram score with $p = 2$ have the highest discrimination heuristic mostly correspond to two scenarios:

(i) Around the financial crisis in 2008, the differences of the distribution forecasts become easier to distinguish. This is because models with a calibration window of 2,000 observations are not heavily affected by the abnormal values during the crisis in contrast to models with a calibration window of only 250 observations. Hence, the distribution forecasts may deviate more strongly between the competing models and scoring rules may assign larger relative distances between the scores of misspecified models and those of the DGP.

(ii) Similarly, the use of FQ-AL as DGP also increases the relative distances between the scores of the models. The Factor Quantile model produces a much sharper forecast than that of alternative models and is therefore easily identifiable as DGP.

In those two cases, all scoring rules manage to clearly identify the DGP from misspecified models, so the even larger relative distance between the scores of the variogram score with $p = 2$ has no additional benefit. Simultaneously, the scoring rule suffers from erroneous rankings, despite having high discrimination heuristics in some settings. These issues show that the discrimination heuristic should only be considered as an indicator for the goodness of scoring rules, but by itself is inadequate to quantify their discrimination ability. A large heuristic of a scoring rule may not imply more robust or less erroneous rankings. Therefore, a high discrimination heuristic between the models is not useful unless it is accompanied with a low error rate, i.e. percentage of choosing a misspecified model over the DGP.

Chapter Eight

Summary and Conclusions

About a decade ago, Gneiting (2008) speculated that the start of the 21st century may usher the transition from point forecasts to distribution forecasts. However, despite a surge of studies on probabilistic forecasting at the beginning of the century, they remain rare in comparison to point forecasts in finance and economics. In this doctoral thesis, we contribute to the current literature on probabilistic predictions in two ways. First, we introduce a flexible semi-parametric model for multivariate distribution forecasting that may be applied easily in higher dimensions. Second, we analyse proper multivariate scoring rules with respect to their ability to identify the true data generating process (DGP) in a realistic setting.

Factor Quantile Models and Related Literature

Our semi-parametric Factor Quantile models offer a simple and efficient way to generate predictions in higher dimensions. Marginals are derived from shape-preserving interpolations on quantiles which in turn are estimated from factor model regressions. We then impose a dependency structure through parametric conditional copulas. Further, we introduce two latent versions of our model in Sections 4.3.1 and 4.3.2 that use endogenous principal component analysis to describe the dependent variables with statistical factors:

(i) The first specification FQ-AL uses the last few principal components and captures the relevant information of the conditional quantile forecasts through the intercept of the regression formula, similar to the concept of Jensen's Alpha. This corresponds to the estimation of future quantiles through their expectation.

(ii) An alternative specification FQ-AB applies bootstrap aggregation (bagging) by Breiman (1996) as a variance reduction technique. This version uses a factor model with the first few principal components as predictors to obtain the asymptotic distribution of the sample quantiles. Then, we generate multiple distribution forecasts by sampling quantiles from their asymptotic distribution. The aggregation of these predictions yields the FQ-AB forecast. This distribution has lower variance than a distribution which directly uses the methodology in Section 4.1 on the principal component representation with the first few principal components.[1] FQ-AB utilizes the entire distribution of the conditional quantiles rather than relying solely on their expectations as in FQ-AL.

Due to the use of uncorrelated principal components in our latent versions, these Factor Quantile models are very robust and exhibit no quantile crossing behaviour in our applications.

We illustrate the general concept of Factor Quantile models with the two-factor Capital Asset Pricing Model (CAPM) introduced by Kraus and Litzenberger (1976) on US stock data in a simple bivariate example in Section 4.2. We apply Clayton, Gumbel and Gaussian copulas to model the dependence between the returns of Apple and Procter & Gamble during the period 2000–2018 and discuss how a dependency structure may be targeted from historical data.

Compared to other forecasting methods with quantile regression, Factor Quantile models can be applied flexibly without reliance on external forecasts or predictors. Due to the multi-stage approach that first estimates marginal distributions and then imposes a dependency structure through a copula, our models scale well in settings with high dimensions. This facilitates their use as a general methodology in many data sets.

[1] We show that a naive principal component representation with the first few principal components has too large a variance to generate accurate predictions directly. Hence, without bagging or similar methods, this model should only be used to estimate confidence intervals for the distribution forecasts.

Our contemporaneous regression model with multiple explanatory factors clearly differs from other quantile regression models in the literature:

(i) Alternative studies such as Cenesizoglu and Timmermann (2008), Zhu (2013) and Pedersen (2015) use lagged, one-factor regressions models. Multivariate information is incorporated into the model by combining the quantile forecasts of different predictors with equal weights. However, in Cenesizoglu and Timmermann (2008), only 16% of the predictors are significant at 1%, raising the question whether forecast averaging with equal weightings can yield appropriate estimates of the future quantiles, when forecasts are included that may be based on inadequate factor models. This is further emphasized by the empirical study of Zhu (2013) where only 9% of the factors for stock returns and 30% for bond returns are significant at 1%. In addition, some of the quantiles such as the median have no significant factor at all.

(ii) Other studies such as Manzan (2015), Bunn et al. (2016) and Meligkotsidou et al. (2019) apply a large set of predictors, possibly with LASSO or similar methods for variable selection. These models show that the general concept of forecasting through quantile regression can yield good results in comparison to benchmark models. However, either they are difficult to apply due to their dependence on the availability of large data sets or they require an understanding of the underlying process to select the explanatory variables. The latent versions of Factor Quantile, in contrast, use the high dimension of the forecasting problem to derive statistical factors through principal component analysis. As such, neither predictor selection nor additional data are necessary.

(iii) Our model can be applied in many general settings since it does not impose any strong assumptions. Conversely, the forecasting methodology by Ma and Pohlman (2008) assumes the conditional location of their dependent variable to be constant over the forecasting period. Other studies such as Gaglianone and Lima (2012) and Bunn et al. (2016) rely on externally generated forecasts

which prohibits the use of their models in more general scenarios. The quantile regression of Gaglianone and Lima (2012) which translates external point forecasts of the expectation of the dependent variable to a distribution forecast may be particularly restrictive. It remains unclear why the expectation of a variable should contain information on other parts of its distribution.

Empirical Evaluation of Factor Quantile Models

For the evaluation of forecasting accuracy we compare two versions of our Factor Quantile model against CCC- and DCC-GARCH, using Student-t asymmetric E-GARCH(1,1) marginals, as well as copulas with EDF marginals. Our time series data include three different multivariate systems of daily USD-denominated exchange rates from 1999–2018, the term structure of US interest rates from 1994–2018 and commodity futures indices from 1991–2018. In contrast to other recent literature on forecasting methodologies, our study makes a significant new empirical contribution to applications of proper multivariate scoring rules, since this is the first such analysis applied to multivariate distribution forecasts of financial asset returns.

We assess the accuracy of forecasts using the MCS of Hansen et al. (2011) derived from the (strictly) proper energy score (Székely, 2003), the variogram score (Scheuerer and Hamill, 2015) and the weighted CRPS (Gneiting and Ranjan, 2011). By evaluating over 1.3 million distribution forecasts in total we highlight how both the scores and the superior model sets depend on the asset class and the timing of the sample.

Previous studies of forecasting models with quantile regression usually only include a limited empirical evaluation and suffer from several common weaknesses:

Short out-of-sample periods: Koenker and Bassett (2010), Gaglianone and Lima (2012) and Manzan (2015), for instance, apply an out-of-sample evaluation based on only 48, 77 and 438 observations respectively. The use of such short evaluation periods is especially relevant, since our empirical and simulation

studies show that scoring rules may not yield correct rankings with such low sample sizes.

Weak benchmark models: Manzan (2015) and Meligkotsidou et al. (2019) use an autoregressive process that is encompassed by their quantile model as a benchmark. The higher relative accuracy is therefore expected since the quantile model incorporates strictly more information than the benchmark and does not get penalized for the excess parameters during testing. Similarly, Cenesizoglu and Timmermann (2008) and Gaglianone and Lima (2012) apply symmetric GARCH models on data with monthly or quarterly frequency. These GARCH models cannot reflect the asymmetric properties of the data adequately and may be unsuited as benchmarks for such low frequencies because volatility clustering is typically only present in data with daily or higher frequency.

Improper evaluation: With the exception of Manzan (2015) and Meligkotsidou et al. (2019), most studies do not apply proper scoring rules and limit their evaluation to simple statistics such as the coverage percentage (Bunn et al., 2016; Gaglianone and Lima, 2012). Furthermore, even when proper scoring rules are employed, the results are difficult to interpret. For instance, Manzan (2015) examines several quantiles separately instead of the entire distribution function which leads to 468 test statistics. The large amount of tests accumulates type I errors and further complicates the identification of the most accurate distribution forecast because the best model varies across the quantiles.

Overall, MCS results based on proper univariate and multivariate scoring rules indicate favourable forecasting performance of both Factor Quantile specifications, matching or exceeding the accuracy of more complicated GARCH models and significantly surpassing the accuracy of copula models with EDF marginals:

(i) The univariate results in Section 6.3.1 measure accuracy through weighted CRPS that focuses on the lower tails, upper tails, both tails, centre of the distribution and the entire distribution. The most successful specification of

FQ-AL remains in 51% of the superior sets on average, compared to 32% for CCC-GARCH, 26% for DCC-GARCH and 14% for EDF. A closer examination of the CRPS test statistic shows that the Factor Quantile models are generally either in the superior set or they are the second best model managing to beat all benchmark models but one. In comparison, FQ-AB, which considers the entire distribution of the conditional quantiles rather than focusing only on their expectation, is included in 61% of the sets. This is higher than the inclusion rate of all benchmark models in all data sets.

(ii) The multivariate comparison in Section 6.3.2 is based on the energy score and the variogram score with $p = 0.5, 1, 2$. Both FQ-AL and FQ-AB stay in 75% of the superior sets with at least one specification. The most successful versions are those that apply the Gaussian copula with a 250 calibration window. These remain in 33% of the sets for FQ-AL and in 67% for FQ-AB. This is comparable to the 50% inclusion rate of DCC-GARCH and much stronger performance than CCC-GARCH and all EDF models which are included in 8% of the superior sets at most.

Generally, the best model depends on the data set employed but Factor Quantile specifications maintain good relative accuracy. The strong performance of Factor Quantile models, even with a simple Gaussian copula, relative to multivariate GARCH models is especially notable since the latter take over five times longer to calibrate in our timing experiments and may also exhibit difficulties with parameter optimisation in eight dimensions. For instance, several parameters of both CCC- and DCC-GARCH converge to unrealistic values for live cattle and sugar in the commodities data even with multi-staged calibration methods implemented in the Oxford MFE Toolbox by Sheppard (2013). These issues require manual attention, which prevents full automation of multivariate GARCH models – see Section 6.1 for a detailed discussion.

Our analysis on several sub-periods in Sections 6.3.1 and 6.3.2 shows that our forecasting accuracy results are robust over time but also emphasizes that scoring

rules need long out-of-sample evaluation periods. For instance, a rolling CRPS test statistic changes the rankings over time and may yield results that are specific to the chosen sample period, even for out-of-sample evaluations with 500 observations. For exchange rate returns, the CRPS inclusion rates for CCC-GARCH range from 20% to 55%, for DCC from 23% to 50%, for FQ-AL from 25% to 65% and for FQ-AB from 38% to 73%, depending on the sub-period. This is especially relevant since many studies in the literature only evaluate their models on short periods.

Comparison of Multivariate Scoring Rules

The evaluation in Section 6.3.2 identifies several issues with multivariate scoring rules that arise due to the high degrees of freedom. Rankings may vary depending on the choice of scoring rule and some scoring rules favour models with independent marginals which certainly do not capture the dependency between the assets adequately. We analyse the ability of the energy score and the variogram score with $p = 0.5, 1, 2$ to distinguish the DGP from misspecified models in our simulation study and find significant differences in the discrimination ability of the four scoring rules.

Our simulation design differs from prior studies by applying a realistic data-driven setting with eight possible choices for the DGP and three data sets that are described in Sections 6.1 and 6.2. We evaluate the scoring rules at around 200 different times t which constitute the first date of each quarter in our evaluation period. In contrast, prior studies in the literature only considered Gaussian distributions as DGP and misspecified models.

Using the scores from the simulation study, we then compare the discrimination ability of the scoring rules through:

(i) The rankings from the sample score means based on large or small sample sizes with 5,000 or 100 scores;

(ii) The error rate, which is the probability that the scoring rule ranks an erroneous distribution higher than the true distribution;

(iii) A discrimination heuristic that measures the relative distance of the sample score means between the correct distribution and all of the misspecified ones. A similar heuristic is used by Pinson and Tastu (2013) to analyse the discrimination ability of the energy score when using a Gaussian DGP.

The conclusions from the sample score means show that the variogram scores with $p = 0.5$ and $p = 1$ produce robust rankings of the models over time and are able to differentiate the DGP from the misspecified models. In contrast, the energy score may prefer slightly misspecified models in small samples and the variogram score with $p = 2$ has severe issues that may lead to erroneous rankings even with a very large sample.

Our error rate also assigns the highest discrimination ability to the variogram scores with $p = 0.5$ and $p = 1$. These results further verify the initial findings based on the sample score means and are robust over all data sets and time periods.

We additionally show that the relative distance of the sample score means is not sufficient to quantify the discrimination ability of the scoring rules. The variogram score with $p = 2$ receives very large values for the discrimination heuristic because it yields larger relative distances when the DGP is easily distinguishable from the misspecified models. However, in other scenarios, where the DGP cannot be clearly identified, this scoring rule often produces erroneous rankings. Hence, a large discrimination heuristic may not lead to correct rankings and may therefore be unsuited by itself for the comparison for scoring rules.

In summary, our simulation study emphasizes the need for large out-of-sample periods and recommends the application of multiple scoring rules in practical applications. Particularly the variogram scores with $p = 0.5$ and $p = 1$ showcase high discrimination ability in our simulation study and may yield more accurate model rankings.

Outlook

The Factor Quantile methodology can be applied with any factor model and (if calibration is not an issue) with any copula. For instance, we have illustrated an application to stock returns using the asymmetric CAPM with a Clayton and Gumbel copula. However, for adequate forecasting accuracy in larger dimensional systems we advocate the use of latent principal component factors. The proven forecasting success of such models paves the way for further work on the application of Factor Quantile models using the factor copula model of Oh and Patton (2017) in place of the more general conditional copula (Patton, 2012) which is employed in this thesis.

Moreover, we have limited the evaluation of our methodology to statistical measures only. Because a good forecast should generate a low expected loss in economic decisions (Elliott and Timmermann, 2016), such as allocating portfolio positions or generating trading strategies, further empirical evaluation of our Factor Quantile methodology using operational tests that measure the economic significance would be interesting.

Bibliography

Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2009), 'Pair-copula constructions of multiple dependence', *Insurance: Mathematics and Economics* **44**(2), 182–198.

Akima, H. (1970), 'A new method of interpolation and smooth curve fitting based on local procedures', *Journal of the ACM (JACM)* **17**(4), 589–602.

Alexander, C. (2002), 'Principal component models for generating large GARCH covariance matrices', *Economic Notes* **31**(2), 337–359.

Alexander, C., Kaeck, A. and Sumawong, A. (2019), 'A parsimonious parametric model for generating margin requirements for futures', *European Journal of Operational Research* **273**(1), 31–43.

Almeida, C., Ardison, K., Kubudi, D., Simonsen, A. and Vicente, J. (2017), 'Forecasting bond yields with segmented term structure models', *Journal of Financial Econometrics* **16**(1), 1–33.

Amin, G. S. and Kat, H. M. (2003), 'Hedge fund performance 1990–2000: Do the "money machines" really add value?', *Journal of Financial and Quantitative Analysis* **38**(2), 251–274.

Amisano, G. and Giacomini, R. (2007), 'Comparing density forecasts via weighted likelihood ratio tests', *Journal of Business & Economic Statistics* **25**(2), 177–190.

Andersen, T., Bollerslev, T., Christoffersen, P. and Diebold, F. (2006), Volatility and correlation forecasting, *in* 'Handbook of Economic Forecasting', Vol. 1, pp. 777–878.

Ando, T. and Tsay, R. S. (2011), 'Quantile regression models with factor-augmented predictors and information criterion', *The Econometrics Journal* **14**(1), 1–24.

Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006), 'Quantile regression under misspecification, with an application to the US wage structure', *Econometrica: Journal of the Econometric Society* **74**(2), 539–563.

Ardia, D. (2008), 'Bayesian estimation of a Markov-switching threshold asymmetric GARCH model with Student-t innovations', *The Econometrics Journal* **12**(1), 105–126.

Asai, M. (2006), 'Comparison of MCMC methods for estimating GARCH models', *Journal of the Japan Statistical Society* **36**(2), 199–212.

Avramidis, P. and Pasiouras, F. (2015), 'Calculating systemic risk capital: A factor model approach', *Journal of Financial Stability* **16**, 138–150.

Bai, X., Russell, J. R. and Tiao, G. C. (2003), 'Kurtosis of GARCH and stochastic volatility models with non-normal innovations', *Journal of Econometrics* **114**(2), 349–360.

Bali, T., Heidari, M. and Wu, L. (2009), 'Predictability of interest rates and interest-rate portfolios', *Journal of Business & Economic Statistics* **27**(4), 517–527.

Bank of International Settlements (2016), Triennial Central Bank survey: Foreign exchange turnover in April 2016, Technical report.

Bao, Y., Lee, T.-H. and Saltoğlu, B. (2007), 'Comparing density forecast models', *Journal of Forecasting* **26**(3), 203–225.

Bauwens, L. and Laurent, S. (2005), 'A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models', *Journal of Business & Economic Statistics* **23**(3), 346–354.

Bauwens, L., Laurent, S. and Rombouts, J. V. (2006), 'Multivariate GARCH models: A survey', *Journal of Applied Econometrics* **21**(1), 79–109.

Belloni, A. and Chernozhukov, V. (2011), 'L1-penalized quantile regression in high-dimensional sparse models', *The Annals of Statistics* **39**(1), 82–130.

Bickel, J. E. (2007), 'Some comparisons among quadratic, spherical, and logarithmic scoring rules', *Decision Analysis* **4**(2), 49–65.

Birge, J. R. (2007), Optimization methods in dynamic portfolio management, *in* 'Handbooks in Operations Research and Management Science', Vol. 15, pp. 845–865.

Bloomberg (2017), The Bloomberg commodity index family: Index methodology, Technical report.

Boero, G., Smith, J. and Wallis, K. F. (2011), 'Scoring rules and survey density forecasts', *International Journal of Forecasting* **27**(2), 379–393.

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of Econometrics* **31**(3), 307–327.

Bollerslev, T. (1987), 'A conditionally heteroskedastic time series model for speculative prices and rates of return', *Review of Economics and Statistics* **69**(3), 542–547.

Bollerslev, T. (1990), 'Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model', *Review of Economics and statistics* **72**(3), 498–505.

Bollerslev, T., Engle, R. F. and Wooldridge, J. M. (1988), 'A capital asset pricing model with time-varying covariances', *Journal of Political Economy* **96**(1), 116–131.

Bollerslev, T. and Wooldridge, J. M. (1992), 'Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances', *Econometric Reviews* **11**(2), 143–172.

Bradley, R. C. (2005), 'Basic properties of strong mixing conditions. A survey and some open questions', *Probability Surveys* **2**, 107–144.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Brockwell, A. E. (2007), 'Universal residuals: A multivariate transformation', *Statistics & Probability Letters* **77**(14), 1473–1478.

Brooks, C., Burke, S. P. and Persand, G. (2003), 'Multivariate GARCH models: Software choice and estimation issues', *Journal of Applied Econometrics* **18**(6), 725–734.

Buja, A., Stuetzle, W. and Shen, Y. (2005), Loss functions for binary class probability estimation and classification: Structure and applications.

Bunn, D., Andresen, A., Chen, D. and Westgaard, S. (2016), 'Analysis and forecasting of electricity price risks with quantile factor models', *Energy Journal* **37**(1), 101–122.

Cai, Y. (2010), 'Multivariate quantile function models', *Statistica Sinica* **20**, 481–496.

Cajigas, J.-P. and Urga, G. (2006), Dynamic conditional correlation models with asymmetric multivariate Laplace innovations.

Capistrán, C. and Timmermann, A. (2009), 'Forecast combination with entry and exit of experts', *Journal of Business & Economic Statistics* **27**(4), 428–440.

Cappiello, L., Engle, R. F. and Sheppard, K. (2006), 'Asymmetric dynamics in the correlations of global equity and bond returns', *Journal of Financial Econometrics* **4**(4), 537–572.

Cenesizoglu, T. and Timmermann, A. G. (2008), Is the distribution of stock returns predictable?

Chakraborty, B. (2001), 'On affine equivariant multivariate quantiles', *Annals of the Institute of Statistical Mathematics* **53**(2), 380–403.

Chakraborty, B. (2003), 'On multivariate quantile regression', *Journal of Statistical Planning and Inference* **110**(1–2), 109–132.

Chaudhuri, P. (1996), 'On a geometric notion of quantiles for multivariate data', *Journal of the American Statistical Association* **91**, 862–872.

Chavas, J.-P. (2018), 'On multivariate quantile regression analysis', *Statistical Methods & Applications* **27**(3), 365–384.

Chen, S.-L., Jackson, J. D., Kim, H. and Resiandini, P. (2014), 'What drives commodity prices?', *American Journal of Agricultural Economics* **96**(5), 1455–1468.

Chen, X. and Fan, Y. (2006*a*), 'Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification', *Journal of Econometrics* **135**(1–2), 125–154.

Chen, X. and Fan, Y. (2006*b*), 'Estimation of copula-based semiparametric time series models', *Journal of Econometrics* **130**(2), 307–335.

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2010), 'Quantile and probability curves without crossing', *Econometrica: Journal of the Econometric Society* **78**(3), 1093–1125.

Chernozhukov, V., Fernández-Val, I. and Melly, B. (2013), 'Inference on counterfactual distributions', *Econometrica: Journal of the Econometric Society* **81**(6), 2205–2268.

Chernozhukov, V. and Umantsev, L. (2001), 'Conditional value-at-risk: Aspects of modeling and estimation', *Empirical Economics* **26**(1), 271–292.

Cheung, Y.-W., Chinn, M. D., Pascual, A. G. and Zhang, Y. (2018), 'Exchange rate prediction redux: New models, new data, new currencies', *Journal of International Money and Finance* .

Chicheportiche, R. and Bouchaud, J.-P. (2012), 'The joint distribution of stock returns is not elliptical', *International Journal of Theoretical and Applied Finance* **15**(03), 1250019.

Chou, R., Yen, T.-J. and Yen, Y.-M. (2017), 'Risk evaluations with robust approximate factor models', *Journal of Banking and Finance* **82**, 244–264.

Christoffersen, P. F. (1998), 'Evaluating interval forecasts', *International Economic Review* **39**(4), 841–862.

Clements, M., Galvao, A. and Kim, J. (2008), 'Quantile forecasts of daily exchange rate returns from forecasts of realized volatility', *Journal of Empirical Finance* **15**, 729–750.

Connor, G. (1995), 'The three types of factor models: A comparison of their explanatory power', *Financial Analysts Journal* **51**(3), 42–46.

Connor, G., Hagmann, M. and Linton, O. (2012), 'Efficient semiparametric estimation of the fama-french model and extensions', *Econometrica: Journal of the Econometric Society* **80**(2), 713–754.

Connor, G. and Korajczyk, R. (1993), 'A test for the number of factors in an approximate factor model', *The Journal of Finance* **48**(4), 1263–1291.

Coroneo, L., Giannone, D. and Modugno, M. (2016), 'Unspanned macroeconomic factors in the yield curve', *Journal of Business & Economic Statistics* **34**(3), 472–485.

Dawid, P. A. (1984), 'Statistical theory: The prequential approach', *Journal of the Royal Statistical Society: Series A (General)* pp. 278–292.

Dawid, P. A. and Sebastiani, P. (1999), 'Coherent dispersion criteria for optimal experimental design', *Annals of Statistics* pp. 65–81.

de Finetti, B. (1975), *Theory of Probability*, Wiley Series in Probability and Statistics.

Diebold, F. X. (2015), 'Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests', *Journal of Business & Economic Statistics* **33**(1), 1–1.

Diebold, F. X., Gunther, T. A. and S, T. A. (1998), 'Evaluating density forecasts, with applications to financial risk management', *International Economic Review* **39**, 863–883.

Diebold, F. X. and Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business & Economic Statistics* **13**(3), 253–263.

Diks, C., Panchenko, V., Sokolinskiy, O. and van Dijk, D. (2014), 'Comparing the accuracy of multivariate density forecasts in selected regions of the copula support', *Journal of Economic Dynamics and Control* **48**, 79–94.

Diks, C., Panchenko, V. and Van Dijk, D. (2010), 'Out-of-sample comparison of copula specifications in multivariate density forecasts', *Journal of Economic Dynamics and Control* **34**(9), 1596–1609.

Dittmar, R. F. (2002), 'Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns', *The Journal of Finance* **57**(1), 369–403.

Duan, J.-C. and Miao, W. (2016), 'Default correlations and large-portfolio credit analysis', *Journal of Business & Economic Statistics* **34**(4), 536–546.

Duffie, D. and Pan, J. (1997), 'An overview of Value at Risk', *Journal of Derivatives* **4**(3), 7–49.

Elliott, G., Gargano, A. and Timmermann, A. (2013), 'Complete subset regressions', *Journal of Econometrics* **177**(2), 357–373.

Elliott, G. and Timmermann, A. (2008), 'Economic forecasting', *Journal of Economic Literature* **46**(1), 3–56.

Elliott, G. and Timmermann, A. (2016), 'Forecasting in Economics and Finance', *Annual Review of Economics* **8**, 81–110.

Ellis, T. and McLain, D. (1977), 'Algorithm 514: A new method of cubic curve fitting using local data', *ACM Transactions on Mathematical Software (TOMS)* **3**(2), 175–179.

Embrechts, P., McNeil, E. and Straumann, D. (1999), 'Correlation: Pitfalls and alternatives', *RISK Magazine* pp. 69–71.

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation', *Econometrica: Journal of the Econometric Society* pp. 987–1007.

Engle, R. F. (2001), 'GARCH 101: The use of ARCH/GARCH models in applied econometrics', *Journal of Economic Perspectives* **15**(4), 157–168.

Engle, R. F. (2002), 'Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models', *Journal of Business & Economic Statistics* **20**(3), 339–350.

Engle, R. F. (2009), *Anticipating correlations: A new paradigm for risk management*, Princeton University Press.

Engle, R. F. and Kroner, K. F. (1995), 'Multivariate simultaneous generalized ARCH', *Econometric Theory* **11**(1), 122–150.

Engle, R. F. and Manganelli, S. (2004), 'CAViaR: Conditional autoregressive value at risk by regression quantiles', *Journal of Business & Economic Statistics* **22**(4), 367–381.

Engle, R. F. and Sheppard, K. (2008), Evaluating the specification of covariance models for large portfolios.

Fama, E. and French, K. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* **33**(1), 3–56.

Feldmann, K., Scheuerer, M. and Thorarinsdottir, T. L. (2015), 'Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression', *Monthly Weather Review* **143**(3), 955–971.

Fenske, N., Kneib, T. and Hothorn, T. (2011), 'Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression', *Journal of the American Statistical Association* **106**(494), 494–510.

Fermanian, J.-D. and Wegkamp, M. H. (2012), 'Time-dependent copulas', *Journal of Multivariate Analysis* **110**, 19–29.

Foresi, S. and Peracchi, F. (1995), 'The conditional distribution of excess returns: An empirical analysis', *Journal of the American Statistical Association* **90**(430), 451–466.

Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics* pp. 1189–1232.

Fritsch, F. N. and Carlson, R. E. (1980), 'Monotone piecewise cubic interpolation', *SIAM Journal on Numerical Analysis* **17**(2), 238–246.

Gaglianone, W. P. and Lima, L. R. (2012), 'Constructing density forecasts from quantile regressions', *Journal of Money, Credit and Banking* **44**(8), 1589–1607.

Gaglianone, W. P., Lima, L. R., Linton, O. and Smith, D. R. (2011), 'Evaluating value-at-risk models via quantile regression', *Journal of Business & Economic Statistics* **29**(1), 150–160.

Garratt, A., Lee, K., Pesaran, M. H. and Shin, Y. (2003), 'Forecast uncertainties in macroeconomic modeling: An application to the UK economy', *Journal of the American Statistical Association* **98**(464), 829–838.

Gebetsberger, M., Messner, J. W., Mayr, G. J. and Zeileis, A. (2018), 'Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood', *Monthly Weather Review* **146**(12), 4323–4338.

Giacomini, R. and White, H. (2006), 'Tests of conditional predictive ability', *Econometrica: Journal of the Econometric Society* **74**(6), 1545–1578.

Gilchrist, W. (2000), *Statistical modelling with quantile functions*, Chapman and Hall/CRC.

Gneiting, T. (2008), 'Probabilistic forecasting', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**(2), 319–321.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007), 'Probabilistic forecasts, calibration and sharpness', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(2), 243–268.

Gneiting, T. and Katzfuss, M. (2014), 'Probabilistic forecasting', *Annual Review of Statistics and Its Application* **1**, 125–151.

Gneiting, T. and Raftery, A. E. (2005), 'Weather forecasting with ensemble methods', *Science* **310**(5746), 248–249.

Gneiting, T. and Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Gneiting, T. and Ranjan, R. (2011), 'Comparing density forecasts using threshold-and quantile-weighted scoring rules', *Journal of Business & Economic Statistics* **29**(3), 411–422.

Gneiting, T. and Ranjan, R. (2013), 'Combining predictive distributions', *Electronic Journal of Statistics* **7**, 1747–1782.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008), 'Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds', *Test* **17**(2), 211.

Goetzmann, W., Ingersoll, J., Spiegel, M. and Welch, I. (2007), 'Portfolio performance manipulation and manipulation-proof performance measures', *The Review of Financial Studies* **20**(5), 1503–1546.

Goyal, A. and Welch, I. (2003), 'Predicting the equity premium with dividend ratios', *Management Science* **49**(5), 639–654.

Granger, C. W. J. and Pesaran, H. M. (2000), A decision theoretic approach to forecast evaluation, *in* 'Statistics and Finance: An interface', World Scientific, pp. 261–278.

Greenaway-McGrevy, R., Mark, N. C., Sul, D. and Wu, J.-L. (2018), 'Identifying exchange rate common factors', *International Economic Review* **59**(4), 2193–2218.

Grimit, E. P., Gneiting, T., Berrocal, V. and Johnson, N. A. (2006), 'The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification', *Quarterly Journal of the Royal Meteorological Society* **132**(621C), 2925–2942.

Groen, J. J., Paap, R. and Ravazzolo, F. (2013), 'Real-time inflation forecasting in a changing world', *Journal of Business & Economic Statistics* **31**(1), 29–44.

Haas, M., Mittnik, S. and Paolella, M. S. (2004), 'Mixed normal conditional heteroskedasticity', *Journal of Financial Econometrics* **2**(2), 211–250.

Hallin, M., Paindaveine, D. and Šiman, M. (2010), 'Multivariate quantiles and multiple-output regression quantiles: From L1 optimization to halfspace depth', *Annals of Statistics* **38**(2), 635–669.

Hamill, T. M. (2001), 'Interpretation of rank histograms for verifying ensemble forecasts', *Monthly Weather Review* **129**(3), 550–560.

Hansen, P. R. (2005), 'A test for superior predictive ability', *Journal of Business & Economic Statistics* **23**(4), 365–380.

Hansen, P. R. and Lunde, A. (2005), 'A forecast comparison of volatility models: Does anything beat a GARCH(1, 1)?', *Journal of Applied Econometrics* **20**(7), 873–889.

Hansen, P. R., Lunde, A. and Nason, J. M. (2011), 'The model confidence set', *Econometrica: Journal of the Econometric Society* **79**(2), 453–497.

Harvey, C. R. and Siddique, A. (2000), 'Conditional skewness in asset pricing tests', *The Journal of Finance* **55**(3), 1263–1295.

Harvey, D., Leybourne, S. and Newbold, P. (1997), 'Testing the equality of prediction mean squared errors', *International Journal of Forecasting* **13**(2), 281–291.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009), *The elements of statistical learning: Data mining, inference, and prediction*, Springer.

Haugom, E., Ray, R., Ullrich, C., Veka, S. and Westgaard, S. (2016), 'A parsimonious quantile regression model to forecast day-ahead value-at-risk', *Finance Research Letters* **16**, 196–207.

He, X. (1997), 'Quantile curves without crossing', *The American Statistician* **51**(2), 186–192.

Hering, C., Hofert, M., Mai, J.-F. and Scherer, M. (2010), 'Constructing hierarchical Archimedean copulas with Lévy subordinators', *Journal of Multivariate Analysis* **101**(6), 1428–1433.

Hofert, M. and Scherer, M. (2011), 'CDO pricing with nested Archimedean copulas', *Quantitative Finance* **11**(5), 775–787.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology* **24**, 417–441 and 498–520.

Hua, J. and Manzan, S. (2013), 'Forecasting the return distribution using high-frequency volatility measures', *Journal of Banking & Finance* **37**(11), 4381–4403.

Jarrow, R. and Protter, P. (2013), 'Positive alphas, abnormal performance, and illusory arbitrage', *Mathematical Finance* **23**(1), 39–56.

Jensen, M. C. (1968), 'The performance of mutual funds in the period 1945–1964', *The Journal of Finance* **23**(2), 389–416.

Johnstone, D. J., Jose, V. R. R. and Winkler, R. L. (2011), 'Tailored scoring rules for probabilities', *Decision Analysis* **8**(4), 256–268.

Jolliffe, I. T. (1986), *Principal component analysis and factor analysis*, Springer-Verlag New York.

Jolliffe, I. T. and Stephenson, D. B. (2003), *A Practitioner's Guide in Atmospheric Science*, Vol. eds.

Jones, C. M. (1994), 'Expectiles and m-quantiles are quantiles', *Statistics & Probability Letters* **20**(2), 149–153.

Jordan, A., Krüger, F. and Lerch, S. (2017), Evaluating probabilistic forecasts with the R package scoringRules.

Judd, K., Smith, L. A. and Weisheimer, A. (2007), 'How good is an ensemble at capturing truth? Using bounding boxes for forecast evaluation', *Quarterly Journal of the Royal Meteorological Society* **133**(626), 1309–1325.

Karlsson, S. (2013), Forecasting with Bayesian vector autoregression, *in* 'Handbook of Economic Forecasting', Vol. 2, Elsevier, pp. 791–897.

Kavtaradze, L. and Mokhtari, M. (2018), 'Factor models and time-varying parameter framework for forecasting exchange rates and inflation: A survey', *Journal of Economic Surveys* **32**(2), 302–334.

Keune, J., Ohlwein, C. and Hense, A. (2014), 'Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts'.

Kilian, L. and Taylor, M. P. (2003), 'Why is it so difficult to beat the random walk forecast of exchange rates?', *Journal of International Economics* **60**(1), 85–107.

Koenker, R. (2004), 'Quantile regression for longitudinal data', *Journal of Multivariate Analysis* **91**(1), 74–89.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press.

Koenker, R. (2011), 'Additive models for quantile regression: Model selection and confidence bandaids', *Brazilian Journal of Probability and Statistics* **25**(3), 239–262.

Koenker, R. and Bassett, G. (1982), 'Robust tests for heteroscedasticity based on regression quantiles', *Econometrica: Journal of the Econometric Society* pp. 43–61.

Koenker, R. and Bassett, G. (2010), 'March madness, quantile regression bracketology, and the Hayek hypothesis', *Journal of Business & Economic Statistics* **28**(1), 26–35.

Koenker, R. and Bassett Jr, G. (1978), 'Regression quantiles', *Econometrica: Journal of the Econometric Society* pp. 33–50.

Koenker, R. and Leorato, S. (2015), Distribution vs. quantile regression.

Koenker, R. W. and d'Orey, V. (1987), 'Computing regression quantiles', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **36**(3), 383–393.

Koenker, R. and Xiao, Z. (2002), 'Inference on the quantile regression process', *Econometrica: Journal of the Econometric Society* **70**(4), 1583–1612.

Koenker, R. and Xiao, Z. (2006), 'Quantile autoregression', *Journal of the American Statistical Association* **101**, 980–990.

Koltchinskii, V. (1997), 'M-estimation, convexity and quantiles', *The Annals of Statistics* pp. 435–477.

Kraus, A. and Litzenberger, R. H. (1976), 'Skewness preference and the valuation of risk assets', *The Journal of Finance* **31**(4), 1085–1100.

Kupiec, P. (1995), 'Techniques for verifying the accuracy of risk measurement models', *The Journal of Derivatives* **3**(2).

Laio, F. and Tamea, S. (2007), 'Verification tools for probabilistic forecasts of continuous hydrological variables', *Hydrology and Earth System Sciences Discussions* **11**(4), 1267–1277.

Laurent, S., Rombouts, J. and Violante, F. (2012), 'On the forecasting accuracy of multivariate GARCH models', *Journal of Applied Econometrics* **27**(6), 934–955.

Leeb, H. and Pötscher, B. M. (2003), 'The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations', *Econometric Theory* **19**(1), 100–142.

Lima, L. R. and Meng, F. (2017), 'Out-of-sample return predictability: A quantile combination approach', *Journal of Applied Econometrics* **32**(4), 877–895.

Ljung, G. M. and Box, G. E. (1978), 'On a measure of lack of fit in time series models', *Biometrika* **65**(2), 297–303.

Ma, L. and Pohlman, L. (2008), 'Return forecasts and optimal portfolio construction: A quantile regression approach', *The European Journal of Finance* **14**(5), 409–425.

Machete, R. L. (2013), 'Contrasting probabilistic scoring rules', *Journal of Statistical Planning and Inference* **143**(10), 1781–1790.

Maciejowska, K., Nowotarski, J. and Weron, R. (2016), 'Probabilistic forecasting of electricity spot prices using factor quantile regression averaging', *International Journal of Forecasting* **32**(3), 957–965.

Mandelbrot, B. B. (1963), 'The variation of certain speculative prices', *Journal of Business* **36**, 394–419.

Manzan, S. (2015), 'Forecasting the distribution of economic variables in a data-rich environment', *Journal of Business & Economic Statistics* **33**(1), 144–164.

Matheson, J. E. and Winkler, R. L. (1976), 'Scoring rules for continuous probability distributions', *Management Science* **22**(10), 1087–1096.

Meligkotsidou, L., Panopoulou, E., Vrontos, I. D. and Vrontos, S. D. (2019), 'Quantile forecast combinations in realised volatility prediction', *Journal of the Operational Research Society* pp. 1–14.

Merkle, E. C. and Steyvers, M. (2013), 'Choosing a strictly proper scoring rule', *Decision Analysis* **10**(4), 292–304.

Min, A. and Czado, C. (2010), 'Bayesian inference for multivariate copulas using pair-copula constructions', *Journal of Financial Econometrics* **8**(4), 511–546.

Nelsen, R. B. (2006), *An introduction to copulas*, Springer Science & Business Media.

Nelson, D. B. (1991), 'Conditional heteroskedasticity in asset returns: A new approach', *Econometrica: Journal of the Econometric Society* pp. 347–370.

Newey, W. K. and Powell, J. L. (1987), 'Asymmetric least squares estimation and testing', *Econometrica: Journal of the Econometric Society* pp. 819–847.

Nowotarski, J. and Weron, R. (2015), 'Computing electricity spot price prediction intervals using quantile regression and forecast averaging', *Computational Statistics* **30**(3), 791–803.

Oh, D. H. and Patton, A. J. (2017), 'Modeling dependence in high dimensions with factor copulas', *Journal of Business & Economic Statistics* **35**(1), 139–154.

Palmer, T. N. (2002), 'The economic value of ensemble forecasts as a tool for risk assessment: From days to decades', *Quarterly Journal of the Royal Meteorological Society* **128**(581), 747–774.

Panagiotelis, A. and Smith, M. (2008), 'Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions', *International Journal of Forecasting* **24**(4), 710–727.

Passow, E. (1974), 'Piecewise monotone spline interpolation', *Journal of Approximation Theory* **12**(3), 240–241.

Patton, A. J. (2006), 'Modelling asymmetric exchange rate dependence', *International Economic Review* **47**(2), 527–556.

Patton, A. J. (2009), Copula-based models for financial time series, *in* 'Handbook of Financial Time Series', Springer, pp. 767–785.

Patton, A. J. (2012), 'A review of copula models for economic time series', *Journal of Multivariate Analysis* **110**, 4–18.

Patton, A. J. (2013), Copula methods for forecasting multivariate time series, *in* 'Handbook of Economic Forecasting', Vol. 2, pp. 899–960.

Pearson, K. (1901), 'On lines and planes of closest fit to systems of points in space', *Philosophical Magazine* **2**(11), 559–572.

Pedersen, T. Q. (2015), 'Predictable return distributions', *Journal of Forecasting* **34**(2), 114–132.

Pelagatti, M. M. (2004), Dynamic conditional correlation with elliptical distributions.

Pierdzioch, C., Risse, M. and Rohloff, S. (2016), 'A quantile-boosting approach to forecasting gold returns', *The North American Journal of Economics and Finance* **35**, 38–55.

Pinson, P. and Girard, R. (2012), 'Evaluating the quality of scenarios of short-term wind power generation', *Applied Energy* **96**, 12–20.

Pinson, P. and Tastu, J. (2013), Discrimination ability of the energy score, Technical report.

Ravazzolo, F. and Vahey, S. P. (2014), 'Forecast densities for economic aggregates from disaggregate ensembles', *Studies in Nonlinear Dynamics & Econometrics* **18**(4), 367–381.

Rémillard, B. (2010), Goodness-of-fit tests for copulas of multivariate time series.

Resta, M. (2012), 'Portfolio optimization: New challenges and perspectives', *Recent Patents on Computer Science* **5**(1), 59–65.

Romano, J. P. and Wolf, M. (2005), 'Stepwise multiple testing as formalized data snooping', *Econometrica: Journal of the Econometric Society* **73**(4), 1237–1282.

Rosenblatt, M. (1952), 'Remarks on a multivariate transform', *Annals of Mathematical Statistics* **23**, 470–472.

Ross, S. A. (1976), 'The arbitrage theory of capital asset pricing', *Journal of Economic Theory* **13**(3), 341–360.

Rothe, C. (2012), 'Partial distributional policy effects', *Econometrica: Journal of the Econometric Society* **80**(5), 2269–2301.

Scheuerer, M. and Hamill, T. M. (2015), 'Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities', *Monthly Weather Review* **143**(4), 1321–1334.

Selten, R. (1998), 'Axiomatic characterization of the quadratic scoring rule', *Experimental Economics* **1**(1), 43–61.

Sheppard, K. (2013), 'Oxford MFE toolbox'. Accessed: 2017-02-07.

Silvennoinen, A. and Terasvirta, T. (2009), Multivariate GARCH models, *in* 'Handbook of Financial Time Series', Springer, pp. 201–229.

Staël von Holstein, C.-A. S. (1970), 'Measurement of subjective probability', *Acta Psychologica* **34**, 146–159.

Steen, M., Westgaard, S. and Gjolberg, O. (2015), 'Commodity value-at-risk modeling: Comparing RiskMetrics, historic simulation and quantile regression', *Journal of Risk Model Validation* **9**(2), 49–78.

Stephenson, D. B. and Dolas-Reyes, F. J. (2000), 'Statistical methods for interpreting Monte Carlo ensemble forecasts', *Tellus A: Dynamic Meteorology and Oceanography* **52**(3), 300–322.

Stock, J. H. and Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**(460), 1167–1179.

Székely, G. J. (2003), 'E-statistics: The energy of statistical samples', *Bowling Green State University, Department of Mathematics and Statistics Technical Report* **3**(05), 1–18.

Taylor, J. W. (1999), 'A quantile regression approach to estimating the distribution of multiperiod returns', *The Journal of Derivatives* **7**(1), 64–78.

Taylor, J. W. (2007), 'Forecasting daily supermarket sales using exponentially weighted quantile regression', *European Journal of Operational Research* **178**(1), 154–167.

Taylor, J. W. (2008), 'Using exponentially weighted quantile regression to estimate value-at-risk and expected shortfall', *Journal of Financial Econometrics* **6**(3), 382–406.

Teräsvirta, T. (2009), An introduction to univariate GARCH models, *in* 'Handbook of Financial Time Series', Springer, pp. 17–42.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the LASSO', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1), 267–288.

Timmermann, A. (2000), 'Density forecasting in Economics and Finance', *Journal of Forecasting* **19**(4), 231–234.

Toda, M. (1963), Measurement of subjective probability distributions, Technical report.

Tsui, A. K. and Yu, Q. (1999), 'Constant conditional correlation in a bivariate GARCH model: Evidence from the stock markets of China', *Mathematics and Computers in Simulation* **48**(4–6), 503–509.

Tu, A. and Chen, C.-H. (2018), 'A factor-based approach of bond portfolio value-at-risk: The informational roles of macroeconomic and financial stress factors', *Journal of Empirical Finance* **45**, 243–268.

Tukey, J. W. (1974), Mathematics and the picturing of of data, *in* 'Proceedings of the International Congress of Mathematicians', Vol. 2, pp. 523–531.

Virbickaite, A., Ausín, M. C. and Galeano, P. (2015), 'Bayesian inference methods for univariate and multivariate GARCH models: A survey', *Journal of Economic Surveys* **29**(1), 76–96.

Virbickaitė, A., Ausín, M. C. and Galeano, P. (2016), 'A Bayesian non-parametric approach to asymmetric dynamic conditional correlation model with application to portfolio selection', *Computational Statistics & Data Analysis* **100**, 814–829.

Wellmann, D. and Trück, S. (2018), 'Factors of the term structure of sovereign yield spreads', *Journal of International Money and Finance* **81**, 56–75.

White, H. (2000), 'A reality check for data snooping', *Econometrica: Journal of the Econometric Society* **68**(5), 1097–1126.

Winkler, R. L. (1971), 'Probabilistic prediction: Some experimental results', *Journal of the American Statistical Association* **66**(336), 675–685.

Winkler, R. L. (1977), Rewarding expertise in probability assessment, *in* 'Decision Making and Change in Human Affairs', Springer, pp. 127–140.

Winkler, R. L. (1996), 'Scoring rules and the evaluation of probabilities', *Test* **5**(1), 1–60.

Zakamulin, V. (2015), 'A test of covariance-matrix forecasting methods', *Journal of Portfolio Management* **41**(3), 97–108.

Zhang, Y. and Nadarajah, S. (2018), 'A review of backtesting for value-at-risk', *Communications in Statistics - Theory and Methods* **47**(15), 3616–3639.

Zhu, M. (2013), 'Return distribution predictability and its implications for portfolio selection', *International Review of Economics & Finance* **27**, 209–223.

Zolotko, M. and Okhrin, O. (2014), 'Modelling the general dependence between commodity forward curves', *Energy Economics* **43**, 284–296.

Appendix A

Model Confidence Set Tables

Table A1: MCS p-values for FQ-AL: Right-tail weighted CRPS

Model	GARCH		EDF		FQ-AL	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.01	1.00**	0.00	0.01	0.00	0.00
CAD	1.00**	0.35**	0.00	0.00	0.00	0.00
CHF	0.07	1.00**	0.00	0.01	0.00	0.00
EUR	0.00	0.00	0.00	0.00	1.00**	0.00
GBP	0.00	1.00**	0.00	0.00	0.00	0.00
JPY	0.37**	0.06	0.00	0.00	1.00**	0.00
NZD	1.00**	0.00	0.00	0.02	0.00	0.00
SEK	0.34**	0.00	0.00	0.00	1.00**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.00	0.00	0.00
1 year	0.00	0.00	1.00**	0.00	0.00	0.00
2 year	0.00	0.00	0.53**	0.00	1.00**	0.00
3 year	0.00	0.00	0.00	0.00	1.00**	0.00
5 year	1.00**	0.00	0.00	0.00	0.38**	0.00
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.11*	0.24*	0.00	0.00	1.00**	0.00
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.01	0.00	1.00**	0.00	0.00	0.01
Live Cattle	0.14*	0.43**	1.00**	0.00	0.00	0.03
Natural Gas	0.00	1.00**	0.00	0.00	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	1.00**	0.00	0.00	0.00	0.00	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

Table A2: MCS p-values for FQ-AL: Left-tail weighted CRPS

Model	GARCH		EDF		FQ-AL	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.86**	0.80**	0.00	0.00	1.00**	0.00
CAD	1.00**	0.00	0.00	0.00	0.00	0.00
CHF	0.00	0.69**	0.00	0.00	1.00**	0.00
EUR	0.53**	0.00	0.00	0.00	0.38**	1.00**
GBP	0.00	1.00**	0.00	0.00	0.00	0.00
JPY	0.00	0.00	0.00	0.00	1.00**	0.00
NZD	1.00**	0.00	0.00	0.00	0.00	0.00
SEK	0.74**	0.00	0.00	0.00	1.00**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.09	0.00	0.09
1 year	0.00	0.00	1.00**	0.00	0.00	0.00
2 year	0.00	0.00	1.00**	0.00	0.54**	0.00
3 year	0.00	0.00	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.00	1.00**	0.00
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.00	0.37**	0.00	0.00	1.00**	0.00
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.00	0.00	0.11*	1.00**	0.00	0.00
Live Cattle	0.00	1.00**	0.00	0.00	0.00	0.00
Natural Gas	0.00	1.00**	0.00	0.00	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	0.00	0.00	0.00	0.00	1.00**	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

Table A3: MCS p-values for FQ-AL: Both-tails weighted CRPS

Model	GARCH		EDF		FQ-AL	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.01	1.00**	0.01	0.03	0.01	0.01
CAD	1.00**	0.00	0.00	0.09	0.00	0.00
CHF	0.04	1.00**	0.00	0.00	0.00	0.00
EUR	0.00	0.00	0.00	0.00	1.00**	0.00
GBP	0.00	1.00**	0.00	0.00	0.00	0.00
JPY	1.00**	0.00	0.00	0.00	0.51**	0.00
NZD	1.00**	0.00	0.00	0.00	0.00	0.00
SEK	0.35**	0.00	0.00	0.00	1.00**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.00	0.00	0.00
1 year	0.00	0.00	1.00**	0.00	0.00	0.00
2 year	0.00	0.00	1.00**	0.00	0.80**	0.00
3 year	0.00	0.00	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.00	1.00**	0.00
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.00	1.00**	0.00	0.07	0.00	0.07
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.01	0.00	1.00**	0.00	0.00	0.00
Live Cattle	0.00	1.00**	0.00	0.00	0.00	0.00
Natural Gas	0.00	1.00**	0.00	0.00	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	0.45**	0.00	0.00	0.00	1.00**	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

Table A4: MCS p-values for FQ-AL: Centre weighted CRPS

Model	GARCH		EDF		FQ-AL	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.00	0.21*	0.00	0.02	1.00**	0.00
CAD	1.00**	0.00	0.00	0.08	0.00	0.00
CHF	0.01	0.37**	0.00	0.00	1.00**	0.00
EUR	0.72**	0.00	0.00	0.00	0.35**	1.00**
GBP	0.00	1.00**	0.00	0.00	0.00	0.00
JPY	1.00**	0.00	0.00	0.00	0.20*	0.00
NZD	1.00**	0.00	0.00	0.00	0.00	0.00
SEK	0.65**	0.00	0.00	0.00	1.00**	0.00
Interest rate changes						
6 month	0.15*	0.00	1.00**	0.00	0.00	0.00
1 year	0.00	0.00	1.00**	0.00	0.00	0.00
2 year	0.00	0.00	0.81**	0.00	1.00**	0.00
3 year	0.00	0.00	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.00	1.00**	0.00
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.37**	0.37**	0.00	0.00	1.00**	0.00
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.01	0.00	1.00**	0.00	0.00	0.00
Live Cattle	0.00	1.00**	0.00	0.00	0.00	0.00
Natural Gas	0.00	1.00**	0.00	0.00	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	1.00**	0.00	0.00	0.00	0.77**	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

Table A5: MCS p-values for FQ-AB: Right-tail weighted CRPS

Model	GARCH		EDF		FQ-AB	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.01	1.00**	0.01	0.01	0.00	0.01
CAD	1.00**	0.35**	0.00	0.00	0.00	0.08
CHF	0.01	1.00**	0.00	0.01	0.00	0.01
EUR	0.00	0.02	0.00	0.00	1.00**	0.00
GBP	0.00	1.00**	0.00	0.00	0.00	0.00
JPY	0.74**	0.02	0.00	0.00	1.00**	0.00
NZD	0.28**	0.00	0.00	0.24*	1.00**	0.00
SEK	1.00**	0.00	0.00	0.00	0.92**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.00	0.00	0.00
1 year	0.14*	0.00	1.00**	0.00	0.00	0.00
2 year	0.07	0.00	0.00	0.00	1.00**	0.00
3 year	0.00	0.00	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.00	1.00**	0.13
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.47**	0.47**	0.00	0.00	1.00**	0.00
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.01	0.01	1.00**	0.00	0.26**	0.00
Live Cattle	0.56**	1.00**	0.00	0.02	0.35**	0.00
Natural Gas	0.00	1.00**	0.00	0.00	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	1.00**	0.01	0.00	0.00	0.00	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

Table A6: MCS p-values for FQ-AB: Left-tail weighted CRPS

Model	GARCH		EDF		FQ-AB	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	1.00**	0.79**	0.00	0.02	0.79**	0.02
CAD	1.00**	0.00	0.00	0.00	0.00	0.00
CHF	0.24*	1.00**	0.00	0.07	0.14	0.14*
EUR	0.00	0.03	0.00	0.00	1.00**	0.00
GBP	0.24*	1.00**	0.00	0.00	0.00	0.00
JPY	0.04	0.04	0.02	0.04	1.00**	0.04
NZD	1.00**	0.00	0.00	0.00	0.00	0.00
SEK	1.00**	0.00	0.00	0.00	0.55**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.09	0.03	0.09
1 year	0.06	0.00	1.00**	0.00	0.00	0.00
2 year	0.00	0.00	1.00**	0.00	0.36**	0.00
3 year	0.00	0.01	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.00	1.00**	0.00
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.00	0.79**	0.00	0.00	1.00**	0.00
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.19*	0.03	1.00**	0.00	0.30**	0.00
Live Cattle	0.00	1.00**	0.00	0.00	0.00	0.00
Natural Gas	0.00	1.00**	0.01	0.01	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	0.00	0.00	0.00	0.00	1.00**	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

Table A7: MCS p-values for FQ-AB: Both-tails weighted CRPS

Model	GARCH		EDF		FQ-AB	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.01	1.00**	0.01	0.02	0.01	0.03
CAD	1.00**	0.00	0.00	0.00	0.00	0.09
CHF	0.00	1.00**	0.00	0.00	0.00	0.01
EUR	0.00	0.00	0.00	0.00	1.00**	0.00
GBP	0.00	1.00**	0.00	0.00	0.00	0.00
JPY	1.00**	0.02	0.01	0.01	0.52**	0.02
NZD	1.00**	0.00	0.00	0.00	0.00	0.00
SEK	0.80**	0.00	0.00	0.00	1.00**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.00	0.00	0.00
1 year	0.09	0.00	1.00**	0.00	0.00	0.00
2 year	0.00	0.00	1.00**	0.00	0.53**	0.00
3 year	0.00	0.02	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.01	1.00**	0.01
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.00	1.00**	0.00	0.07	0.00	0.07
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.17*	0.02	0.00	0.00	1.00**	0.01
Live Cattle	0.00	1.00**	0.00	0.00	0.00	0.00
Natural Gas	0.00	1.00**	0.00	0.00	0.00	0.00
Soybean	0.00	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	1.00**	0.00	0.00	0.00	0.94**	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

Table A8: MCS p-values for FQ-AB: Centre weighted CRPS

Model	GARCH		EDF		FQ-AB	
	CCC	DCC	250	2000	250	2000
Exchange rate returns						
AUD	0.01	0.99**	0.01	0.01	1.00**	0.02
CAD	0.84**	0.00	0.00	0.19*	1.00**	0.00
CHF	0.02	1.00**	0.00	0.00	0.00	0.02
EUR	0.00	0.00	0.00	0.00	1.00**	0.00
GBP	0.00	0.31**	0.00	0.00	1.00**	0.00
JPY	0.00	0.01	0.00	0.00	1.00**	0.00
NZD	0.48**	0.00	0.00	0.00	1.00**	0.00
SEK	0.00	0.15*	0.00	0.00	1.00**	0.00
Interest rate changes						
6 month	1.00**	0.00	0.00	0.00	0.00	0.16*
1 year	0.08	0.00	1.00**	0.00	0.00	0.00
2 year	0.04	0.00	0.00	0.00	1.00**	0.00
3 year	0.00	0.00	0.00	0.00	1.00**	0.00
5 year	0.00	0.00	0.00	0.00	1.00**	0.00
7 year	0.00	0.00	0.00	0.00	1.00**	0.00
10 year	0.00	0.00	0.00	0.00	1.00**	0.00
20 year	0.00	0.00	0.00	0.00	1.00**	0.00
Commodity index returns						
Copper	0.00	0.00	0.00	0.06	1.00**	0.06
Corn	1.00**	0.00	0.00	0.00	0.00	0.00
Gold	0.02	0.01	1.00**	0.00	0.31**	0.00
Live Cattle	0.07	1.00**	0.00	0.00	0.34**	0.00
Natural Gas	0.00	0.70**	0.00	0.00	1.00**	0.00
Soybean	0.02	1.00**	0.00	0.00	0.00	0.00
Sugar	0.00	0.00	0.00	0.00	1.00**	0.00
WTI Oil	0.00	0.01	0.00	0.00	1.00**	0.00

The MCS p-values are obtained using the entire out-of-sample data listed in Table 6.3. Models with p-values greater than 0.25 or 0.10 are marked with ** or *. These models are included in the superior set with $\alpha = 0.25$ or $\alpha = 0.10$ respectively. The best model is highlighted in blue.

Appendix B

Table Extracts

Table B1: Bloomberg Commodity Index 2019 target weights

Commodity	Target weights	Commodity	Target weights
Energy		*Industrial Metals*	
Natural Gas	8.26%	Copper	7.32%
WTI Crude Oil	7.66%	Aluminium	4.41%
Brent Crude Oil	7.34%	Zinc	3.21%
Low Sulfur Gas Oil	2.63%	Nickel	2.71%
RBOB Gasoline	2.28%	*Precious Metals*	
ULS Diesel	2.16%	Gold	12.24%
Grains		Silver	3.89%
Soybeans	6.03%	*Softs*	
Corn	5.89%	Sugar	3.15%
Soybean Meal	3.44%	Coffee	2.48%
Wheat	3.14%	Cotton	1.42%
Soybean Oil	3.10%	*Livestock*	
HRW Wheat	1.29%	Live Cattle	4.09%
		Lean Hogs	1.85%

The table presents the 2019 target weights which determine the composition of the Bloomberg Commodity Index (BCOM). They are determined in accordance with the rules described in Bloomberg (2017) and were announced by Bloomberg on 31 October 2018. The eight commodity sub-indices we consider in Chapters 6 and 7 are highlighted in blue, make up for 54.64% of the Bloomberg Commodity Index and cover all six sectors.

Table B2: Currency distribution of global foreign exchange market turnover

Currency	Symbol	% of daily trades
United States Dollar	USD	87.6%
Euro	EUR	31.4%
Japanese Yen	JPY	21.6%
Pound Sterling	GBP	12.8%
Australian Dollar	AUD	6.9%
Canadian Dollar	CAD	5.1%
Swiss Franc	CHF	4.8%
Renminbi	CNY	4.0%
Swedish Krona	SEK	2.2%
New Zealand Dollar	NZD	2.1%

The table lists the percent of daily trades (bought or sold) announced by Bank of International Settlements (2016). The eight currencies we consider in Chapters 6 and 7 are highlighted in blue and make up for 86.9% of the global foreign exchange market turnover.